

AD-A135 408

DEVELOPMENT OF STATISTICAL TECHNIQUES TO BETTER UTILIZE 1/1

DATA CHARACTERIZE. (U) VERSAR INC SPRINGFIELD VA

A 5 GLEIT 24 OCT 83 784 AFOSR-TR-83-0951

UNCLASSIFIED

F49620-82-C-0079

F/G 12/1

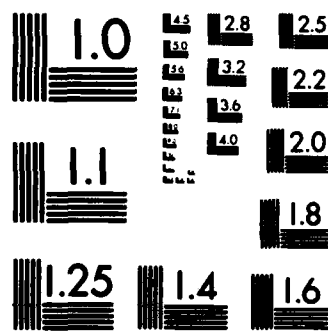
NL

END

FILED

1 84

DTIC



MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A

10

AD-A135408

Final Report

DEVELOPMENT OF STATISTICAL TECHNIQUES TO BETTER  
UTILIZE DATA CHARACTERIZED BY BEING BELOW  
INSTRUMENT DETECTION THRESHOLDS AND BY SMALL  
SAMPLE SIZE

Contract F49620-82-C-0079

Alan S. Gleit  
Versar, Inc.  
P.O. Box 1549  
6850 Versar Center  
Springfield, VA 22151

DTIC  
DEC 6 1983  
A

October 24, 1983


DTIC FILE COPY

Approved for public release;  
distribution unlimited.

80 12 00 000

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER <b>AFOSR-TR- 83 - 0951</b>	2. GOVT ACCESSION NO. <b>AD-A135408</b>	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) <b>DEVELOPMENT OF STATISTICAL TECHNIQUES TO BETTER UTILIZE DATA CHARACTERIZED BY BEING BELOW INSTRUMENT DETECTION THRESHOLDS AND BY SMALL SAMPLE SIZE</b>		5. TYPE OF REPORT & PERIOD COVERED <b>Final Report July 1, 1982 to Aug.31,1983</b>
7. AUTHOR(s) <b>Alan S. Gleit</b>		6. PERFORMING ORG. REPORT NUMBER <b>784</b>
9. PERFORMING ORGANIZATION NAME AND ADDRESS <b>Versar Inc. P.O. Box 1549, 6850 Versar Center Springfield, VA 22151</b>		8. CONTRACT OR GRANT NUMBER(s) <b>F49620-82-C-0079</b>
11. CONTROLLING OFFICE NAME AND ADDRESS <b>AFOSR / NM Directorate of Mathematical and Information Sciences Bolling AFB, DC 20332</b>		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS <b>PE61102F 2304/A5</b>
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		12. REPORT DATE <b>24</b> <b>October</b> , 1983
		13. NUMBER OF PAGES <b>74</b>
		15. SECURITY CLASS. (of this report) <b>unclassified</b>
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) <b>Approved for public release. Distribution unlimited.</b>		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) <b>Estimation, Below Detection Limits, Type I Censoring, Environmental Data, Small data sets</b>		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) <b>Estimation of parametric families for small data sets where a significant portion of the data lay below fixed instrument detection thresholds was investigated. Thus the number of data points was random (an example of Type I censoring). Both analytic and simulation procedures were utilized. In particular, maximum likelihood techniques, order statistic techniques, truncation techniques, fill-in with constants, and fill-in with expected values of the missing points were investigated. For exponential data, truncation seemed most appropriate while for normal and log-normal data, fill-in with expected</b>		

values (modified to correct for conditioning on the number of data points) was best. The criteria for selection was the total square error.



### Objective of Research Effort

Among the problems encountered in attempting to analyze data from actual experiments are (1) a significant portion of the data points often fall below the instrument detection thresholds and (2) insufficient data are available to form the population size necessary to validate conclusions reached by standard statistical techniques. Versar addressed these deficiencies via this research effort to provide the Air Force with better techniques to evaluate experiments yielding data thus characterized.

When measuring environmental phenomena, the measuring devices/procedures used are often unable to detect low concentrations. Thus, concentrations below certain threshold levels are not measurable. Standard "detection limits" are set by various agencies for various phenomena for various types of measuring devices. Measured values below these limits are reported as "below detection limit" and are thus not available for statistical analysis. (Sometimes values below these limits are available, but their accuracy is greatly in doubt.) Consequently, the statistician often has a very basic problem facing him: how does he analyze data sets which contain a reasonable percentage of "below detection limit" entries? This problem is exacerbated by the usual problem of small sample size. As an example, suppose we have taken eight samples of air near a chemical warehouse in order to see if there are leaks. Concentrations below 0.7 parts per billion, say, are below the reliability of the measurement procedure. Of the eight samples, suppose five are below the detection limit while the other three are measured to have concentrations of 1, 2, and 5 parts per billion. How do we find the average concentration?

The dual problems of small sample size and sub-detection limit data can often be encountered by statisticians working on Air Force problems. Examples are:

- o The determination of the "hardening" characteristics of AWACS and other Air Force systems against nuclear explosions. The tests to simulate segments of a nuclear environment are expensive and provide relatively few data points in small portions of the radiation spectrum. A significant portion of these data could be "real," but could be below the detection limits of the instrumentation used.

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)  
NOTICE OF TECHNICAL INFORMATION (DTIC)  
This technical report has been reviewed and is  
approved for release under E.O. 11652 (AFR 19)-12.  
Distribution is unlimited.  
MATTHEW J. KRAMER  
Chief, Technical Information Division

- o The determination of the effects of Chemical, Biological, and Radiation (CBR) warfare agents against Air Force systems and personnel. Again, due to the nature and costs of tests, data would be relatively sparse. Much of these data, particularly those representing "leakage" and other unintentioned side-effects, could be below detection instrument thresholds, but would be useful in evaluating the agent effects, particularly in the regime of low dosages over a long exposure period. The CBR requirement will undoubtedly receive more emphasis after the President's recent announcement to resume the development of these agents.
- o The exposure of Air Force personnel to fumes released by fueling operations. Some fuels contain trace amounts of toxic substances (indications exist that synthetic fuels, which may receive increasing Air Force use, contain larger concentrations of such substances than do conventional fuels). The exposure concentrations, and their effects over time, to persons continually involved in refueling operations need to be further assessed. Many of these concentrations are frequently below detection equipment threshold.

We studied the problem of "below detection limit" data coupled with small size both theoretically and via computer simulations. We suppose that we are given  $N$  data points,  $p$  of which are "below detection limit"  $L$  and  $N-p$  of which have reported values larger than  $L$ . We suppose that the distribution for the underlying stochastic process is known to belong to a fixed family of distributions depending on an unknown parameter  $\theta$ . We wish to estimate  $\theta$ . Among the techniques we used were maximum likelihood techniques, order statistic techniques, truncation techniques, and fill-in with constants or expected values procedures.



A/

### Significant Findings of Research Effort

The findings of our research effort are embodied in three manuscripts:

1. Estimation of the mean for small data sets of left-censored exponential data (Appendix A).
2. Estimation of the normal population parameters by order statistics given a singly censored Type I sample (Appendix B).
3. Estimation of the parameters for small data sets of left censored normal and lognormal data (Appendix C).

In the first, we investigated exponential data characterized by the dual problems of small sample size with several values reported as "smaller than the limit L." We proposed several estimators for the mean. In particular, we investigated:

1. maximum likelihood estimator (MLE)
2. modified MLE, which removes the conditioning of the MLE due to knowledge of  $p$ , the number of censored points
3. best linear invariant estimator
4. best linear unbiased estimator
5. fill-in with constants
6. modified fill-in with constants, which removes the conditioning on  $p$
7. fill-in with expected values, which is equivalent to the MLE
8. truncation.

To evaluate the performance of these procedures we performed a simulation. Selecting (by scaling)  $L=1$ , we tested  $\theta=1/3, 2/3, 1, 2, 3$ , and 5 for sample sizes  $N=5, 10$ , and 15. Using 20,000 data sets for each of the 18 cases, we found that the truncation method was the best while the modified MLE was a close second. Here we used the square error as our criterion for selecting techniques.

The second manuscript deals with estimating the parameters of the censored normal distribution via order statistic techniques. Data from a censored normal has been analyzed many times before using order



statistics. However, all these previous studies used Type II censoring: the  $p$  smallest observations are missing where  $p$  is fixed a priori. Type I censored data (observations below a fixed value are missing) are usually analyzed by Type II methods. We provide Type I estimators; however, the algorithms fail to converge often enough to make the method practical.

The third manuscript deals with estimating the parameters of the censored normal distribution by other than order statistic techniques. In particular we investigated:

1. maximum likelihood estimator (MLE)
2. modified MLE, which removes the conditioning on  $p$
3. fill-in with constants
4. modified fill-in with constants, which removes the conditioning on  $p$
5. fill-in with expected values of the missing data
6. modified fill-in with expected values, which removes the conditioning on  $p$
7. truncation.

To evaluate the performance of these procedures, we performed a simulation. Selecting (by scaling)  $L=1$ , we tested

$$\mu=0.67, \sigma=.2, .3$$

$$\mu=1.00, \sigma=.1, .2, .3$$

$$\mu=1.33, \sigma=.2, .3$$

for sample sizes  $N=5, 10$ , and  $15$ . Using  $50,000$  data sets for each of the  $21$  cases, we found that the modified fill-in with expected values was the best while the fill-in with expected values was only marginally worse. The former had smaller bias but larger variance leading to a slight improvement (in general) of the total squared error.

### Written Manuscripts

The three manuscripts mentioned above will be submitted to appropriate journals for review and possible publications.

### Presentation of Results

We presented the results of this research at the following meetings:

1. Operations Research Society of America, spring 1983 national meeting in Chicago, "Small sample, below detection limit exponential data"
2. Operations Research Society of America, fall 1983 national meeting in Orlando, "Normal data and detection limits."

We provided copies of our manuscripts to several individuals including:

1. Dr. John Beauchamp, Oak Ridge National Laboratory, Tennessee
2. Dr. David Payton, Air Force Weapons Laboratory, Kirtland AFB, New Mexico.

We discussed the material of this study informally with participants in the workshop on reliability held at the University of North Carolina - Charlotte in June 1983.

APPENDIX A

ESTIMATION OF THE MEAN FOR SMALL DATA SETS OF  
LEFT CENSORED EXPONENTIAL DATA

Alan Gleit\*  
Versar Inc.  
6850 Versar Center  
P. O. Box 1549  
Springfield, Virginia 22151

\* This work was sponsored by the Air Force Office of Scientific Research  
under contract F49620-82-C-0079.

# ABSTRACT

We study exponential data characterized by the dual problems of small sample size with several values reported as "smaller than the limit L." In particular, we propose several estimators and report the results of a simulation.

Key words: Below detection limit; Reliability

## INTRODUCTION

In reliability theory, the exponential distribution plays an important role. It usually leads to simple formulas for the quantities of interest. In this way it may provide a "first approximation" to the real-life situation. Indeed, quite often it leads to useful bounds for these quantities. So an investigation of a problem in reliability theory may begin with a discussion of the exponential distribution.

We have in mind the problem of estimating mean shelf-life for objects placed in inventory. Often, to decide whether the object is still operable (or edible or . . .), an expensive or destructive or time-consuming test need be performed. Hence, the number of objects for our experiment is small and so asymptotic or large-sample-size results are inapplicable. Further, since testing requires money and time, one usually does not continuously monitor for failures from the moment that the objects are placed in inventory. Consequently, some of the units might have failed prior to our testing. Thus our data will be characterized by (1) small samples and (2) reported values for failures if above some limit  $L$ , but only "below  $L$ " for those that failed very quickly. Our problem is to estimate the mean of such a data set if we assume the underlying distribution is exponential.

The problem that we address below is an example of censoring. In general, censoring means that observations at one or both extremes are not available. Our problem is equivalent to "left censoring"; life testing usually involves "right censoring", i.e., the largest values are not available. Two types of life censoring have received much attention. Type I occurs when the test is terminated at a specified time before all the items have failed; Type II occurs when the test is terminated at a particular failure. In Type I censoring, the number of failures as well as the failure times are random variables. This of course makes Type I

censoring far more complicated. Consequently Type II methods have often been applied to Type I data with the hope that the bias is not appreciable. Our problem is analogous to Type II censoring since the number of units, say  $p$ , with failures "below  $L$ " is a random variable.

## SECTION 1. THE ESTIMATORS

Suppose we have  $N$  identical units on test with time-to-failure exponentially distributed with parameter  $\theta$ , i.e., time-to-failure has probability density

$$f_{\theta}(x) = \frac{1}{\theta} e^{-x/\theta} \quad x > 0. \quad (1)$$

We assume  $0 < p < N$  values lie below the (known) limit  $L$ . Thus we are given data

$$\{x_1, \dots, x_K, p \text{ values below } L\} \quad (2)$$

where we have taken

$$K = N - p.$$

We are asked to find the parameter  $\theta$ . In this section we shall investigate several techniques to estimate  $\theta$ .

### I. Maximum Likelihood Estimator (MLE)

For our data (2) the likelihood function is given by

$$F_{\theta}(L)^p \prod_{i=1}^K f_{\theta}(x_i)$$

where  $f_{\theta}$  is given in (1) and

$$F_{\theta}(x) = \int_0^x f_{\theta}(t) dt = 1 - e^{-x/\theta}. \quad (3)$$

Substituting (1) and (3) into the likelihood function and taking logarithms yields

$$\log \text{likelihood} = p \log (1 - e^{-L/\theta}) - K \log \theta - \sum x/\theta. \quad (4)$$

Maximizing the expression (4) yields the conditional MLE  $\theta^*$ .

Proposition 1. The conditional MLE  $\theta^*$  exists and is the unique root of

$$\theta^* = \frac{pL}{K(e^{L/\theta^*} - 1)} + \frac{\sum x}{K}. \quad (5)$$

It is consistent, asymptotically efficient, and has asymptotic variance

$$N^{-1} \left[ \frac{L^2}{\exp(L/\theta) - 1} + \theta^2 (\exp(-L/\theta) - \exp(-x_1/\theta)) + \theta^2 \sum_{j=2}^K (\exp(-x_{j-1}/\theta) - \exp(-x_j/\theta)) \right]^{-1}.$$

Proof Follows from Kulldorff (1961, Theorems 11.1 and 11.2).

The estimator  $\theta^*$  is biased. By using a simulation (see section 2 below) we may view the extent of the bias.

Examples. Let  $L=1$ ,  $\theta=2$ ,  $N=15$ . Suppose  $\sum x = K(L+\theta)$ , its expected value (see Prop. A.4(f)).

Then  $p = 4$  gives  $\theta^* = 2.32$

$p = 5$                       2.15

$p = 7$                       1.81.



## II. Modified MLE

The MLE  $\theta^*$  is extremely biased for small samples. The problem is that  $\theta^*$  tries to estimate the average value of all the data. Let

$$A = E(\text{average all data} | p). \quad (6)$$

From Proposition A.4(f), we see that

$$A = \theta + L - \frac{p}{N} \frac{L}{1 - e^{-L/\theta}} \quad (7)$$

which is not  $\theta$ . So we suggest modifying  $\theta^*$  to form the estimator  $\theta_0$  which will satisfy the following implicit formula:

$$\theta^* = \theta_0 + L - \frac{p}{N} \frac{L}{1 - e^{-L/\theta_0}} \quad (8)$$

By using our simulation (see section 2 below), we see that  $\theta_0$  is a very good estimator. Further, if we replace  $\Sigma x$  by its expected value, we see that  $\theta_0$  is close, but not equal, to  $\theta$ . Consequently, it does have some small bias.

Examples. Let  $L=1$ ,  $\theta=2$ ,  $N=15$ . Replacing  $\Sigma x$  by its expected value, we obtain:

$p = 4$	gives $\theta_0 = 2.002$
$p = 5$	2.001
$p = 7$	1.996.

## III. Best Linear Invariant Estimator (BLIE)

We let

$$H = \sum_{i=1}^K C_i x_i \quad (9)$$

be an arbitrary linear estimator. We wish to select those coefficients  $\{C_1, \dots, C_K\}$  which minimize the variance of  $H$  among all invariant  $H$ , i.e. all  $H$  satisfying

$$E(H/\theta) = \text{constant}.$$

From Prop. A.4, we have

$$\begin{aligned}
 E(H) &= \sum C_i E(x_i) \\
 &= \sum C_i (L + E_i \theta) \\
 &= L \sum C_i + \theta \sum C_i E_i.
 \end{aligned} \tag{10}$$

This is a constant times  $\theta$  if and only if  $\sum C_i = 0$ . Hence, invariance of  $H$  is equivalent to

$$\sum C_i = 0. \tag{11}$$

So we wish to minimize  $E(H-\theta)^2$  subject to (11). Now from Prop. A.4 we have

$$\begin{aligned}
 E(H-\theta)^2 &= E(\sum \sum C_i C_j x_i x_j - 2\theta \sum C_i x_i + \theta^2) \\
 &= \sum \sum C_i C_j (D_{ij} \theta^2 + E_i E_j \theta^2 + L E_i \theta + L E_j \theta + \theta^2) \\
 &\quad - 2\theta \sum C_i (L + \sum E_i \theta) + \theta^2 \\
 &= \sum \sum C_i C_j (D_{ij} \theta^2 + E_i E_j \theta^2) - 2\theta \sum C_i E_i + \theta^2
 \end{aligned} \tag{12}$$

where we have used (10). So our problem is to minimize (12) subject to (11).

Using a Lagrange multiplier  $\lambda$ , our problem is to

$$\min L = \min (\sum \sum C_i C_j (D_{ij} + E_i E_j) \theta^2 - 2\theta \sum C_i E_i - 2\lambda \sum C_i). \tag{13}$$

Setting  $\partial L / \partial C_i = 0$ , each  $i$ , yields

$$0 = \sum_j C_j (D_{ij} + E_i E_j) \theta^2 - \theta E_i - \lambda \tag{14}$$

and  $\partial L / \partial \lambda = 0$  yields (11). The solution to our system (11) and (14) is

$$C_1 = -1 + 1/K$$

$$C_i = 1/K, \quad i = 2, \dots, K$$

$$\lambda = -\theta^2/K^2 + L\theta.$$

The verification requires use of Corollary A.3. Hence the BLIE is

$$H^* = -x_1 + \frac{1}{K} \Sigma x.$$

We now have the following result.

Proposition 2. The BLIE is (for  $N-p = K \geq 2$ )

$$H^* = -x_1 + \frac{1}{K} \Sigma x. \quad (15)$$

It satisfies

$$EH^* = \frac{K-1}{K} \theta \quad (16)$$

$$\text{var } H^* = \theta^2/K. \quad (17)$$

Proof. We note that

$$H^* = \frac{K-1}{K} \left( -\frac{K}{K-1} x_1 + \frac{1}{K-1} \Sigma x \right).$$

So from Corollary A.6,  $H^*$  is distributed as  $\frac{\theta}{2K} \chi^2(2(K-1))$ . Hence (16)

follows. To obtain (17) we have

$$\begin{aligned} \text{var } H^* &= E(H^* - \theta)^2 + \text{bias}^2 \\ &= \frac{\theta^2}{4K^2} 4(K-1) + \left( \frac{K-1}{K} - 1 \right)^2 \theta^2 \\ &= \theta^2/K. \end{aligned}$$

#### IV. Best Linear Unbiased Estimator (BLUE)

We let

$$G = \sum_{i=1}^K B_i x_i \quad (18)$$

be an arbitrary linear estimator. We wish to select those coefficients  $\{B_1, \dots, B_K\}$  which minimize the variance of  $G$  among all unbiased  $G$ .

From (10) we have

$$EG = L \Sigma B_i + \theta \Sigma B_i E_i.$$

Since we require  $EG = \theta$ , we require

$$\sum B_i = 0 \quad (19)$$

$$\sum B_i E_i = 1. \quad (20)$$

Hence our problem is  $\min E(G-\theta)^2$  subject to (19) and (20). Now from (12)

$$E(G-\theta)^2 = \sum \sum B_i B_j D_{ij} \theta^2 + 2\theta^2 - 2\theta. \quad (21)$$

Using Lagrange multipliers  $\lambda$  and  $\mu$ , our problem is to

$$\min L = \min (\sum \sum B_i B_j D_{ij} \theta^2 - 2\lambda \sum B_i - 2\mu (\sum B_i E_i - 1)). \quad (22)$$

Setting  $\partial L / \partial B_i = 0$ , each  $i$ , yields

$$0 = \sum B_j D_{ij} \theta^2 - \lambda - \mu E_i, \quad (23)$$

$\partial L / \partial \lambda = 0$  yields (19), and  $\partial L / \partial \mu = 0$  yields (20). The solution to our system (19), (20), and (23) is

$$B_1 = -1$$

$$B_j = 1/(K-1), \quad j=2,3,\dots,K$$

$$\lambda = \theta^2 / K(K-1)$$

$$\mu = -\theta^2 / (K-1).$$

Hence the BLUE is

$$G^* = -\frac{K}{K-1} x_1 + \frac{1}{K-1} \sum x.$$

We have the following result.

Proposition 3. The BLUE is (for  $N-p = K \geq 2$ )

$$G^* = -\frac{K}{K-1} x_1 + \frac{1}{K-1} \sum x. \quad (24)$$

It is distributed as  $\frac{\theta}{2(K-1)} \chi^2 (2(K-1))$ . Consequently,

$$\text{var } G^* = \theta/(K-1). \quad (25)$$

Also the BLIE  $H^*$  satisfies

$$H^* = \frac{K-1}{K} G^*.$$

Proof. See Corollary A.6.

#### V. Fill-in with Constants Approach

Various constants have been suggested as proxies for the data below  $L$ . Pessimists might use zero (i.e. equipment failed immediately) while optimists might argue for  $L$  (i.e. equipment failed at the instant we started checking). Those suggesting some sort of balance might use  $L/2$ . Let us suppose that we use the value  $C$  as a proxy. Then our estimator is

$$\theta^* = \frac{1}{N} (\sum x + pC). \quad (26)$$

This procedure is very easy to use and is easily understood by the statistically non-sophisticated.

Clearly the rule  $\theta^*$  is biased. In fact

$$\begin{aligned} E\theta^* &= \frac{1}{N} (K(L+\theta) + pC) \\ &= \theta + L - \frac{p}{N} (\theta + L - C) \end{aligned} \quad (27)$$

using Proposition A.4(e). Consequently

$$\begin{aligned} \text{var } \theta^* &= \frac{1}{N^2} \text{var } (\sum x) \\ &= \frac{K}{N^2} \theta^2 \end{aligned} \quad (28)$$

from Corollary A.5, a relatively small value since a (sometimes large) part of the data is replaced by a fixed constant. Hence  $\theta^*$  has a very narrow spread about the wrong value!

To improve this technique, we suggest that  $\theta^*$  is trying to estimate

$$\begin{aligned} A &= E(\text{average } |p) \\ &= \theta + L - \frac{P}{N} \frac{L}{1 - e^{-L/\theta}} \end{aligned}$$

from (7). Our suggestion is to modify  $\theta^*$  to form  $\theta_0$  which will satisfy

$$\theta^* = \theta_0 + L - \frac{P}{N} \frac{L}{1 - e^{-L/\theta_0}}. \quad (29)$$

Our simulation (see Section 2 below) shows that  $\theta_0$  is a much improved estimator.

#### VI. Fill-in with Expected Values

Let us fill in the missing data not by constants as in V above but by more appropriate values: their expected values. From Proposition A.4(d) we have

$$E(\text{sum of missing data } |p) = p(\theta - \frac{L}{e^{L/\theta} - 1}). \quad (30)$$

Hence our estimator  $\theta^*$  satisfies the equation

$$\theta^* = \frac{1}{N} [Lx + p(\theta^* - \frac{L}{e^{L/\theta^*} - 1})]. \quad (31)$$

After rearranging, this equation is identical to (5), the equation for the MLE! Consequently, the MLE procedure is equivalent to filling-in the data points "below L" with their conditional expectations. This interpretation adds credence to our suggestion in II that the MLE needs to be adjusted via the procedures outlined there.

## VII. Truncation

Our last technique is very easy to conceptualize: forget that data below  $L$  has been obtained and assume that the distribution of the remaining  $K$  data points are governed by the truncated exponential distribution

$$\begin{aligned} g_{\theta}(x) &= \frac{\frac{1}{\theta} e^{-x/\theta}}{\int_L^{\infty} \frac{1}{\theta} e^{-t/\theta} dt} \\ &= \frac{1}{\theta} e^{(L-x)/\theta} \quad \text{for } x > L. \end{aligned} \quad (32)$$

All "good" estimators (i.e. MLE, BLUE, Minimum Variance Unbiased Estimator) for our truncated distribution (32) are the same:

$$\theta^* = \frac{1}{K} \sum x - L. \quad (33)$$

It has the following properties.

Proposition 4. The truncated estimator  $\theta^*$  is distributed as

$$\frac{\theta}{2K} \chi^2(2K).$$

As such, it is unbiased with variance

$$\text{var } \theta^* = \theta^2/K,$$

which is smaller than that of the BLUE.

Proof. See Corollary A.6(a).

## SECTION 2. THE SIMULATION

In order to evaluate the performance of the estimators based on maximum likelihood procedures (parts I and II of section 1) and on filling-in by constants (part V of section 1), we performed a simulation. Since the exponential distribution has only a scale parameter  $\theta$  to estimate, all the formulas depend only on the ratio  $\theta/L$ . Thus we were free to normalize the simulated data to the case  $L=1$ ,  $\theta=1/3, 2/3, 1, 2, 3$ , and  $5$ . We selected  $N=5, 10$ , and  $15$  as representative small data set sizes.

Using a standard pseudo-standard generator, we simulated 20,000 data sets for each value of  $N$  and  $\theta$ . The data sets were then artificially censored at the cutoff  $L=1$  and passed to the several estimators to "guess" values for  $\theta$ . The data sets were then grouped by  $p$ , the number of missing data values, and averaged. Typical results are included in Tables 1 and 2 below. The tables include the method of truncation (part VII of section 1) as a means to check the simulation since the mean and variance for this method have been theoretically calculated in Proposition 4. We can clearly see that the theoretical values agree quite well with those obtained in the simulation.

(Insert Tables 1,2 about here.)



### CONCLUSION

We have presented above several methods to estimate  $\theta$  based on censored from below data sets. Several are extremely simple, easily calculated and understood by the mathematically unsophisticated. Among these, the method of truncation

$$\theta^* = \frac{1}{N-p} \sum x - L$$

is clearly the best. It is unbiased and has small variance

$$\text{var } \theta^* = \theta^2 / (N-p).$$

Simulated data shows that it performs just about as well as predicted.

For the more sophisticated worker who will use a computer to find an estimator the modified MLE (part II of Section 1) appears to be slightly superior. It is found via a two-step procedure:  $\theta_0$  satisfies

$$\theta_0 + L - P/N \frac{L}{1 - e^{-L/\theta_0}} = \theta^*$$

where  $\theta^*$  satisfies

$$\theta^* = \frac{1}{(N-p)} \sum x - \frac{pL}{(N-p)(e^{L/\theta^*} - 1)}$$

It appears to have little bias with slightly smaller variance than the method of truncation.

## APPENDIX

In this section we investigate the distributions of various random variables associated with our estimators.

Proposition A.1 Let  $Y_1, \dots, Y_K$  be the order statistics for an exponential distribution with parameter  $\theta$ , i.e.  $Y_1, \dots, Y_K$  are a random sample arranged in ascending order  $0 \leq Y_1 \leq Y_2 \leq \dots \leq Y_K$ .

Then

- a.  $Y_1$  has an exponential distribution with parameter  $\theta/K$ .
- b.  $Y_{j+1} - Y_j$  has an exponential distribution with parameter  $\theta/(K-j)$ ,  $j=1, \dots, K-1$
- c.  $\{Y_1, Y_2 - Y_1, \dots, Y_K - Y_{K-1}\}$  are independent.

Corollary A.2 For the situation described in Proposition 1,

- a.  $E(Y_j) = (\frac{1}{K} + \frac{1}{K-1} + \dots + \frac{1}{K-j+1})\theta$
- b.  $\text{Cov}(Y_i, Y_j) = ((\frac{1}{K})^2 + (\frac{1}{K-1})^2 + \dots + (\frac{1}{K-m+1})^2)\theta^2$

where  $m = \min(i, j)$ .

We let

$$E_j^{(K)} = \frac{1}{K} + \dots + \frac{1}{K-j+1}$$

$$D_{ij}^{(K)} = (\frac{1}{K})^2 + \dots + (\frac{1}{K-m+1})^2$$

$$D_j^{(K)} = D_{jj}^{(K)} = \sum_{n=0}^{j-1} (\frac{1}{K-n})^2.$$

Corollary A.3 For the quantities defined above, we have:

- a.  $\sum_{j=1}^K D_{ij}^{(K)} = E_i^{(K)}$
- b.  $\sum_{i=1}^K E_i^{(K)} = K$
- c.  $\sum \sum D_{ij}^{(K)} = K.$

Proof. Suppressing the superscripts, we have the following calculations,

$$\begin{aligned}
 \text{a. } \sum_{j=1}^K D_{ij} &= \sum_{j=1}^i D_{ij} + \sum_{j=i+1}^K D_{ij} \\
 &= \sum_{j=1}^i D_j + \sum_{j=i+1}^K D_i \\
 &= \sum_{j=1}^i D_j + (K-i)D_i \\
 &= \left(\frac{1}{K}\right)^2 + i\left(\frac{1}{K}\right)^2 + \left(\frac{1}{K-1}\right)^2 + \dots + \left[\left(\frac{1}{K}\right)^2 + \left(\frac{1}{K-1}\right)^2 + \dots + \left(\frac{1}{K-i+1}\right)^2\right] + (K-i)D_i \\
 &= i\left(\frac{1}{K}\right)^2 + (i-1)\left(\frac{1}{K-1}\right)^2 + (i-2)\left(\frac{1}{K-2}\right)^2 + \dots + 1\left(\frac{1}{K-i+1}\right)^2 + (K-i)D_i \\
 &= \sum_{n=0}^{i-1} (i-n)\left(\frac{1}{K-n}\right)^2 + (K-i) \sum_{n=0}^{i-1} \left(\frac{1}{K-n}\right)^2 \\
 &= \sum_{n=0}^{i-1} (K-n)\left(\frac{1}{K-n}\right)^2 \\
 &= \sum_{n=0}^{i-1} \frac{1}{K-n} \\
 &= E_i.
 \end{aligned}$$

$$\begin{aligned}
 \text{b. } \sum_{i=1}^K E_i &= \left(\frac{1}{K}\right) + \left(\frac{1}{K} + \frac{1}{K-1}\right) + \dots + \left(\frac{1}{K} + \frac{1}{K-1} + \dots + \frac{1}{1}\right) \\
 &= K\left(\frac{1}{K}\right) + (K-1)\left(\frac{1}{K-1}\right) + \dots + (1)\frac{1}{1} \\
 &= K.
 \end{aligned}$$

We suppose in the remainder of this section that  $0 < L < \infty$  is a fixed, given (known) constant.

Proposition A.4 Let  $Y_1, \dots, Y_N$  be the order statistics from an exponential distribution with parameter  $\theta$ . Suppose  $Y_p < L \leq Y_{p+1}$  and let  $K=N-p$ . Let

$$X_j = Y_{p+j} \quad j=1, \dots, K.$$

Then a.  $\{X_j - L\}$  are the order statistics from an exponential distribution with parameter  $\theta$ .

$$b. E(X_j | p) = E_j^{(K)} \theta + L$$

$$c. \text{cov}(X_i, X_j | p) = D_{ij}^{(K)} \theta^2$$

$$d. E(Y_1 + \dots + Y_p | p) = p \left[ \theta - \frac{L}{e^{L/\theta} - 1} \right]$$

$$e. E(X_1 + \dots + X_K | p) = (N-p)(L+\theta)$$

$$f. E(Y_1 + \dots + Y_N | p) = N(L+\theta) - \frac{pL}{1 - e^{-L/\theta}}.$$

Proof d.  $E(Y_1 + \dots + Y_p | p) = E(\text{sum of } p \text{ independent samples each less than } L)$

$$= pE(X < L)$$

$$= p \left[ \frac{\theta - (\theta + L)e^{-L/\theta}}{1 - e^{-L/\theta}} \right]$$

$$= p \left[ \theta - \frac{L}{e^{L/\theta} - 1} \right].$$

e. follows from Corollary 3(b) and part (b).

f. follows from parts (d) and (e).

Corollary A.5 For the situation described in Proposition A.4, we have

$$\text{var}(\sum X|p) = K\theta^2$$

Proof 
$$\begin{aligned}\text{var}(\sum X|p) &= E((\sum X)^2|p) - (E(\sum X|p))^2 \\ &= E(\sum \sum X_i X_j | p) - (K(L+\theta))^2 \\ &= \sum \sum (D_{ij}\theta^2 + (L+E_i\theta)(L+E_j\theta)) - (K(L+\theta))^2 \\ &= K\theta^2 + (K(L+\theta))^2 - (K(L+\theta))^2\end{aligned}$$

Corollary A.6 For the situation described in Proposition A.4, we have

$$\begin{aligned}\text{a. } & (\frac{1}{K}\sum X - L) \sim \frac{\theta}{2K} \chi^2(2K) \\ \text{b. } & (-\frac{K}{K-1}X_1 + \frac{1}{K-1}\sum X) \sim \frac{\theta}{2(K-1)} \chi^2(2(K-1))\end{aligned}$$

Proof. Let

$$\begin{aligned}S_1 &= K(X_1 - L) \\ S_j &= (K-j+1)(X_j - X_{j-1}) \quad j=2, \dots, K.\end{aligned}$$

Then

$$\begin{aligned}E(S_j) &= (K-j+1)(EX_j - EX_{j-1}) \\ &= (K-j+1)(E_j^{(K)} - E_{j-1}^{(K)}) \\ &= (K-j+1)\theta / (K-j+1) \\ &= \theta.\end{aligned}$$

Thus  $\{S_j/\theta\}$  are i.i.d. exponential 1. Hence  $2S_j/\theta \sim \chi^2(2)$  and so

$$\sum_{j=T}^K S_j \sim \frac{\theta}{2} \chi^2(2(K-T+1))$$

and

$$\frac{1}{K-T+1} \sum_{j=T}^K S_j \sim \frac{\theta}{2(K-T+1)} \chi^2(2(K-T+1)).$$

Part (a) is the case  $T=1$ ; part (b) is the case  $T=2$ .

#### REFERENCES

KULLDORFF, GUNNAR, (1961). Contributions to the Estimation from Grouped and Partially Grouped Samples, New York: John Wiley.

Table 1. Simulation Results for N=5, True Mean = 2/3, L=1<sup>a</sup>

Method	p=1			p=2			p=3			p=4		
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
MLE	1.4231	0.0601	1.1638	0.0548	0.9072	0.0390	0.6268	0.0221	0.6268	0.0221	0.6268	0.0221
Modified MLE	0.6874	0.0867	0.6982	0.1267	0.7234	0.1743	0.7164	0.2857	0.7164	0.2857	0.7164	0.2857
Fillin with 1.0	1.5351	0.0592	1.3930	0.0525	1.2632	0.0352	1.1327	0.0171	1.1327	0.0171	1.1327	0.0171
Fillin with 0.5	1.4351	0.0592	1.1930	0.0525	0.9632	0.0352	0.7327	0.0171	0.7327	0.0171	0.7327	0.0171
Fillin with 0.0	1.3351	0.0592	0.9931	0.0525	0.6632	0.0352	0.3327	0.0171	0.3327	0.0171	0.3327	0.0171
Mod. fillin with 1.0	0.8210	0.0875	1.0458	0.1338	1.4964	0.2014	2.7875	0.4004	2.7875	0.4004	2.7875	0.4004
Mod. fillin with 0.5	0.7014	0.0857	0.7403	0.1238	0.8305	0.1697	1.0343	0.2946	1.0343	0.2946	1.0343	0.2946
Mod. fillin with 0.0	0.5840	0.0832	0.4624	0.1048	0.3251	0.0985	0.1637	0.0632	0.1637	0.0632	0.1637	0.0632
Truncation	0.6688	0.0925	0.6551	0.1457	0.6581	0.2202	0.6634	0.4275	0.6634	0.4275	0.6634	0.4275
Truncation (Theory)	0.6667	0.1111	0.6667	0.1481	0.6667	0.2222	0.6667	0.4444	0.6667	0.4444	0.6667	0.4444

<sup>a</sup> The distribution of the 20,000 samples was: p=0, 10 cases; p=1, 164 cases; p=2, 1341 cases; p=3, 4715 cases; p=4, 8150 cases; p=5, 5620 cases. The overall mean for all data was 0.6658.

Table 2. Simulation Results for  $N=10$ , True Mean = 2,  $t=1^a$

Method	p=1		p=2		p=3		p=4		p=5		p=6		p=7		p=8		p=9	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
MF	2.1173	0.3054	2.4999	2.3247	2.2357	0.2725	1.9928	0.2420	1.7290	0.2100	1.4499	0.1582	1.2079	0.1225	0.9442	0.0932		
Modified MF	2.0354	0.4735	2.0116	0.5015	1.9998	0.5453	2.0199	0.6512	2.0000	0.8024	1.9556	0.9176	1.9913	1.2132	1.9119	1.9100		
Fillin with 1.0	2.8305	0.3046	2.6069	0.3230	2.3914	0.2701	2.2105	0.2381	2.0045	0.2045	1.7065	0.1513	1.6092	0.1137	1.4190	0.0826		
Fillin with 0.5	2.7805	0.3046	2.5069	0.3229	2.2414	0.2699	2.0105	0.2380	1.7545	0.2046	1.4065	0.1513	1.2592	0.1137	1.0190	0.0826		
Fillin with 0.0	2.7305	0.3046	2.4070	0.3228	2.0974	0.2700	1.8105	0.2380	1.5045	0.2045	1.1065	0.1513	0.9092	0.1137	0.6190	0.0826		
Mod. fillin with 1.0	2.0943	0.4727	2.1444	0.4998	2.2282	0.5425	2.3162	0.6473	2.5453	0.7900	2.7665	0.9162	3.2633	1.7208	4.1819	2.0043		
Mod. fillin with 0.5	2.0389	0.4725	2.0202	0.4990	2.0163	0.5402	2.0405	0.6415	2.0570	0.7044	2.0411	0.8060	2.1450	1.1546	2.2045	1.8515		
Mod. fillin with 0.0	1.9835	0.4724	1.8962	0.4981	1.8053	0.5370	1.7241	0.6310	1.5003	0.7532	1.3562	0.7954	1.1551	0.9019	0.8450	1.1281		
Truncation	2.0139	0.4748	2.0087	0.5046	1.9963	0.5509	2.0175	0.6613	2.0091	0.8102	1.9663	0.9454	2.0308	1.2633	2.0950	2.0660		
Truncation (library)	2.0	0.4444	2.0	0.5	2.0	0.5114	2.0	0.6667	2.0	0.8	2.0	1.0	2.0	1.3332	2.0	2.0	2.0	4.0

<sup>a</sup> The distribution of the 20,000 samples was: p=0, 128 cases; p=1, 871 cases; p=2, 2596 cases; p=3, 4654 cases; p=4, 5008 cases; p=5, 3921 cases; p=6, 2170 cases; p=7, 175 cases; p=8, 201 cases; p=9, 25 cases; p=10, 1 case. The overall mean for all data was 2.0040.



APPENDIX B

ESTIMATION OF THE NORMAL POPULATION PARAMETERS BY  
ORDER STATISTICS GIVEN A SINGLY CENSORED TYPE I SAMPLE

Alan Gleit\*  
Versar Inc.  
6850 Versar Center  
P. O. Box 1549  
Springfield, Virginia 22151

\* This work was sponsored by the Air Force Office of Scientific Research  
under contract F49620-82-C-0079.

# ABSTRACT

We construct the Best Linear Unbiased Estimators for the mean and variance given a Type I censored sample from a normal population. Numerical experience with small data sets indicates that our iterative procedure to find the estimators almost never converges.

Key words: Best Linear Unbiased Estimation, BLUE

## INTRODUCTION

The problem of estimating the parameters from a censored normal distribution has been extensively treated in the literature. Two natural censoring mechanisms are: (1) observations below or above a given point may be missing (Type I) and (2) the  $p$  smallest or largest observations of a sample of size  $N$  may be missing (Type II). Type I censoring is more complex since the number of observations  $K = N - p$  is a random variable. Consequently, Type II censoring methodology has been applied to Type I data though the methods are clearly biased.

One widely used method to estimate the parameters of a normal distribution is based on linear combinations of order statistics. For Type II censoring, the known sample elements are arranged in ascending order, i.e.,  $X(1) \leq X(2) \leq \dots \leq X(K)$ , and the method of least squares is applied to get the best linear combination of them. The coefficients provided by these linear estimators are unbiased (if  $K$  is known a priori) with minimal variance. Important contributions to this methodology include Gupta (1952), Sarhan and Greenberg (1956, 1958), Law (1959) and Dixon (1960). Below we extend this methodology to the case of Type I censoring.

## SECTION 1. CONDITIONAL ESTIMATORS

Let

$$X_1 \leq X_2 \leq \dots \leq X_K$$

be the ordered censored sample of size  $K$  out of a complete sample of size  $N$ . We assume that the  $p = N - K$  censored values are known to lie below the censoring value  $L$ . We will first develop the minimal variance unbiased linear estimator (BLUE) conditional on  $p$ . Later we shall remove this conditioning.

Our initial problem is to find

$$G = \sum_j G_j X_j \quad (1)$$

and

$$H = \sum_j H_j X_j \quad (2)$$

with  $E(G) = \mu$ ,  $E(H) = \sigma$ , and minimal variance among such linear estimators. To better formulate one problem, let

$E(i, R, K)$  = expected value of the  $i$ th order statistic from groups of size  $K$  for a standard normal random variable censored from below at  $R$ ,

$cov(i, j, R, K)$  = expected value of the covariance of the  $i$ th and  $j$ th order statistics from groups of size  $K$  for a standard normal random variable censored from below at  $R$ .

Using this notation we have

$$E(X_j) = \mu + \sigma E(j, (L - \mu)/\sigma, K) \quad (3)$$

$$cov(X_i, X_j) = \sigma^2 cov(i, j, (L - \mu)/\sigma, K). \quad (4)$$

Hence

$$EG = \sum_j G_j (\mu + \sigma E(j, (L - \mu)/\sigma, K)) \quad (5)$$

$$var G = \sigma^2 \sum_i \sum_j G_i G_j cov(i, j, (L - \mu)/\sigma, K) \quad (6)$$

with similar formulas for H. Since EG should be  $\mu$ , we obtain from (5):

$$\begin{aligned}\sum_j G_j &= 1 \\ \sum_j G_j E(j, (L-\mu)/\sigma, K) &= 0.\end{aligned}$$

Thus we may formulate our problem as follows.

Proposition 1. The BLUE's for  $\mu$  and  $\sigma$  solve the following problems.

$$\begin{aligned}P1: \min \sum_{ij} G_i G_j \text{cov}(i, j, (L-\mu)/\sigma, K) \\ \text{such that } \sum_j G_j &= 1\end{aligned}\tag{7}$$

$$\sum_j G_j E(j, (L-\mu)/\sigma, K) = 0\tag{8}$$

and

$$\begin{aligned}P2: \min \sum_{ij} H_i H_j \text{cov}(i, j, (L-\mu)/\sigma, K) \\ \text{such that } \sum_j H_j &= 0\end{aligned}\tag{9}$$

$$\sum_j H_j E(j, (L-\mu)/\sigma, K) = 1.\tag{10}$$

To solve P1 and P2, let us first write them using the equivalent Lagrangean formulation:

$$P1': \min \sum_{ij} G_i G_j \text{cov}(i, j, (L-\mu)/\sigma, K) - \alpha_1 (\sum_j G_j - 1) - \beta_1 \sum_j G_j E(j, (L-\mu)/\sigma, K)$$

$$P2': \min \sum_{ij} H_i H_j \text{cov}(i, j, (L-\mu)/\sigma, K) - \alpha_2 \sum_j H_j - \beta_2 (\sum_j H_j E(j, (L-\mu)/\sigma, K) - 1).$$

To write down the solution to P1' and P2', we first introduce some additional notation. Let

$\Sigma(R, K)$  = matrix with  $(i, j)$  coordinate equal to  $\text{cov}(i, j, R, K)$

$E(R, K)$  = vector with  $i$ th coordinate equal to  $E(i, R, K)$

$1$  = vector of length  $K$  whose elements are all the number one.

Finally let

$$M(R, K) = \begin{pmatrix} \Sigma(R, K) & -1' & -E'(R, K) \\ 1 & 0 & 0 \\ E(R, K) & 0 & 0 \end{pmatrix}.$$

Then the solutions to P1' and P2' are

$$\begin{pmatrix} G \\ \alpha_1 \\ \beta_1 \end{pmatrix} = M^{-1}((L-\mu)/\sigma, K) \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad (11)$$

$$\begin{pmatrix} H \\ \alpha_2 \\ \beta_2 \end{pmatrix} = M^{-1}((L-\mu)/\sigma, K) \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad (12)$$

with optimal variances  $\alpha_1$  and  $\beta_2$ .

Unfortunately, the optimal linear combinations  $G$  and  $H$  depend on the unknown parameters  $\mu$  and  $\sigma$ . Hence our estimates for  $\mu$  and  $\sigma$  need to be found by iterative procedures.

**Proposition 2.** The estimates  $\mu^*$  and  $\sigma^*$  provided by the BLUE's satisfy (11), (12), and

$$\begin{aligned} \sum_j G_j x_j &= \mu^* \\ \sum_j H_j x_j &= \sigma^*. \end{aligned}$$

Solutions to these four equations can be found provided extensive tables of  $M^{-1}(R, K)$  are available. If they were, a useful algorithm is fairly straightforward:

1. Let  $R_0 = 0$
2. Let  $G_{I+1}, H_{I+1}$  satisfy

$$\begin{pmatrix} G_{I+1} \\ \alpha_1 \\ \beta_1 \end{pmatrix} = M^{-1}(R_I, K) \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} H_{I+1} \\ \sigma_2 \\ s_2 \end{pmatrix} = M^{-1}(R, K)_I \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

3. Let

$$\mu_{I+1} = \sum G_{I+1,j} X_j$$

$$\sigma_{I+1} = \sum H_{I+1,j} X_j$$

$$R_{I+1} = (L - \mu_{I+1}) / \sigma_{I+1}.$$

4. If  $|\mu_I - \mu_{I+1}|$  and  $|\sigma_I - \sigma_{I+1}|$  are small enough, stop.  
Otherwise, let  $I \leftarrow I+1$  and go to step 2.

## SECTION 2. UNCONDITIONAL ESTIMATORS

We now, after considerable effort, have  $\mu^*$ ,  $\sigma^*$  which clearly depend on  $K$ . To obtain unconditional estimators we need to find the expected value of  $\mu^*$ ,  $\sigma^*$ . For this purpose, we note the following.

### Proposition 3.

$$1. E(x | x \geq L) = \mu + \sigma \frac{\phi((L-\mu)/\sigma)}{1-\Phi((L-\mu)/\sigma)} \quad (13)$$

$$2. E(x | x \leq L) = \mu - \sigma \frac{\phi((L-\mu)/\sigma)}{\Phi((L-\mu)/\sigma)} \quad (14)$$

$$3. E(x^2 | x \geq L) = \mu^2 + \sigma^2 + (L+\mu)\sigma \frac{\phi((L-\mu)/\sigma)}{1-\Phi((L-\mu)/\sigma)} \quad (15)$$

$$4. E(x^2 | x \leq L) = \mu^2 + \sigma^2 - (L+\mu)\sigma \frac{\phi((L-\mu)/\sigma)}{\Phi((L-\mu)/\sigma)} \quad (16)$$

Letting

$$A = \frac{\phi((L-\mu)/\sigma)}{1-\Phi((L-\mu)/\sigma)} \quad (17)$$

and

$$B = \frac{\phi((L-\mu)/\sigma)}{\Phi((L-\mu)/\sigma)} \quad (18)$$

and using (13)-(16) we may easily obtain the following.

### Corollary 4.

$$1. E(\text{sum of all } N \text{ data} | p \text{ missing}) = N\mu + \sigma(KA - pB)$$

$$2. E(\text{sum of squares of all } N \text{ data} | p \text{ missing}) =$$

$$N(\mu^2 + \sigma^2) + (L+\mu)\sigma[KA - pB]$$

$$3. E((\text{sum of all } N \text{ data})^2 | p \text{ missing}) =$$

$$K(K-1)(\mu + \sigma A)^2 + 2pK(\mu + \sigma A)(\mu - \sigma B) + p(p-1)(\mu - \sigma B)^2 + \\ + N(\mu^2 + \sigma^2) + (L+\mu)\sigma(KA - pB).$$



Corollary 5.

1.  $E(\text{average of all } N \text{ data} | p \text{ missing}) = \mu + \sigma(KA - pB)/N$  (19)

2.  $E(S^2 \text{ for all } N \text{ data} | p \text{ missing}) =$   
 $\mu^2 + \sigma^2 + (L + \mu)\sigma(KA - pB)/N - K(K-1)(\mu + \sigma A)^2/N(N-1)$   
 $- 2pK(\mu + \sigma A)(\mu - \sigma B)/N(N-1) - p(p-1)(\mu - \sigma B)^2/N(N-1).$  (20)

Hence the expected values of  $\mu^*$  and  $\sigma^{*2}$  are not  $\mu$  and  $\sigma^2$  but the expressions on the right hand sides of (19) and (20). Another iterative scheme would convert the biased  $\mu^*$  and  $\sigma^{*2}$  to their unbiased counterparts.

### SECTION 3. NUMERICAL EXPERIENCE

To perform the calculations indicated in Section 1, rather extensive tables of  $M^{-1}(R,K)$  would be required. In turn, to find the inverse for the matrix  $M(R,K)$  we would need to find the expected values  $E(R,K)$  and the variance  $\Sigma(R,K)$  for groups of  $K$  order statistics from a standard normal distribution with censor value  $R$ . To find these we generated one million normal variates. We then grouped all those values greater than  $R$  into groups of size  $K$ . The means and covariances were computed as the averages of the values for each such group. Tables were prepared for  $R = -3.0, -2.0(.1)+2.0, +3.0$  and  $K=1(1)15$ . As an example, we show in Tables I, II, III below the output for  $R=.1$ ,  $K=1,2,3,4,5,9$ , and 14. For groups of 14 values from the standard normal distribution all above the value  $R=.1$ , the first ( $J=1$ ) order statistic has expected value 0.1763, the second ( $J=2$ ) order statistic has expected value 0.2549, etc. Further, the variance of the first ( $J=1$ ) order statistic is 0.0053, of the second ( $J=2$ ) order statistics is 0.0103, etc. Also, the covariance of the second ( $J=2$ ) and first ( $I=1$ ) order statistics is 0.0049, etc.

For each possible  $(R,K)$  combination  $M(R,K)$  and thence  $M^{-1}(R,K)$  was found. For values of  $R$  not in our tables, linear interpolation was used.

To test the value of our estimators, we used various combinations of  $\mu$  and  $\sigma$ . By adjusting the range and scale, we chose  $L=1$  and

$$\mu=1.33; \sigma=.2,.3$$

$$\mu=1.00; \sigma=.1,.2,.3$$

$$\mu=.67; \sigma=.2,.3$$

with total sample sizes  $N=5, 10$ , and 15. For each  $(\mu, \sigma, N)$  combination we generated 50,000 samples, censored them at the value  $L=1$ , and tried our algorithm on the resulting data sets. In the algorithm of

Section 1 we let "small" be .01 and stopped if R was ever out of range. Unfortunately, in no  $(\mu, \sigma, K)$  instance did even 10% of the samples converge for  $-3 \leq \frac{1-\mu^*}{\sigma^*} \leq +3$ ; in fact, for only  $\mu=1, \sigma=.1$  did even 5% coverage!

Hence, the methodology described above, though theoretically useful, has little practical value.

TABLE I. Expected values of standard normal variates  
above the value 0.1 in groups of K= 1,2,3,4, and 5

TRUNCATION VALUE = 0.1					
	MEAN OF X	MEAN OF X**2	MEAN OF X**3	MEAN OF X**4	COVARIANCE
K= 1 J= 1 I= 1	0.8616	1.0847	1.7301	3.25E+0	0.3425
SUM OF MEAN:	0.8616				
K= 2 J= 1 I= 1	0.5380	0.4209	0.4256	5.14E-1	0.1315
J= 2 I= 1		0.7357			0.1027
2	1.1768	1.7346	3.0215	5.99E+0	0.3501
SUM OF MEAN:	1.7147				
K= 3 J= 1 I= 1	0.4134	0.2442	0.1887	1.77E-1	0.0733
J= 2 I= 1		0.3893			0.0623
2	0.7911	0.7818	0.9159	1.23E+0	0.1560
J= 3 I= 1		0.6165			0.0502
2		1.2093			0.1255
3	1.3700	2.2069	4.0492	8.27E+0	0.3302
SUM OF MEAN:	2.5745				
K= 4 J= 1 I= 1	0.3438	0.1643	0.1023	7.74E-2	0.0461
J= 2 I= 1		0.2535			0.0408
2	0.6186	0.4759	0.4341	4.53E-1	0.0933
J= 3 I= 1		0.3671			0.0350
2		0.6778			0.0803
3	0.9661	1.0926	1.4049	2.01E+0	0.1595
J= 4 I= 1		0.5472			0.0286
2		0.9979			0.0649
3		1.5845			0.1273
4	1.5085	2.5878	4.9481	1.04E+1	0.3124
SUM OF MEAN:	3.4370				
K= 5 J= 1 I= 1	0.3007	0.1227	0.0647	4.16E-2	0.0323
J= 2 I= 1		0.1841			0.0286
2	0.5172	0.3314	0.2523	2.20E-1	0.0640
J= 3 I= 1		0.2575			0.0257
2		0.4561			0.0573
3	0.7711	0.6970	0.7181	8.26E-1	0.1024
J= 4 I= 1		0.3513			0.0223
2		0.6148			0.0489
3		0.9317			0.0879
4	1.0944	1.3557	1.8653	2.81E+0	0.1581
J= 5 I= 1		0.5038			0.0182
2		0.8747			0.0395
3		1.3156			0.0703
4		1.8963			0.1290
5	1.6150	2.9093	5.7664	1.24E+1	0.3012
SUM OF MEAN:	4.2984				

TABLE II. Expected values of standard normal variates  
above the value 0.1 in groups of K=9

-----						
TRUNCATION VALUE = 0.1						
-----						
		MEAN OF X	MEAN OF X**2	MEAN OF X**3	MEAN OF X**4	COVARIANCE
K= 9	J= 1 I= 1	0.2161	0.0581	0.0194	7.84E-3	0.0115
	J= 2 I= 1		0.0840			0.0108
		2	0.3389	0.1383	3.73E-2	0.0235
	J= 3 I= 1		0.1108			0.0101
		2	0.1797			0.0218
		3	0.4659	0.1544	1.06E-1	0.0345
	J= 4 I= 1		0.1399			0.0095
		2	0.2249			0.0204
		3	0.3130			0.0318
		4	0.6035	0.3075	2.52E-1	0.0453
	J= 5 I= 1		0.1722			0.0088
		2	0.2752			0.0189
		3	0.3819			0.0295
		4	0.4984			0.0419
		5	0.7564	0.5755	5.67E-1	0.0544
	J= 6 I= 1		0.2091			0.0081
		2	0.3328			0.0174
		3	0.4607			0.0271
		4	0.5999			0.0383
		5	0.7580			0.0541
		6	0.9307	1.0251	1.20E+0	0.0749
	J= 7 I= 1		0.2536			0.0072
		2	0.4020			0.0154
		3	0.5557			0.0244
		4	0.7230			0.0348
		5	0.9119			0.0492
		6	1.1293			0.0680
		7	1.4005	1.8372	2.56E+0	0.0994
	J= 8 I= 1		0.3113			0.0062
		2	0.4923			0.0135
		3	0.6796			0.0216
		4	0.8833			0.0308
		5	1.1122			0.0437
		6	1.3751			0.0606
		7	1.7000			0.0890
		8	1.4127	3.4460	5.89E+0	0.1420
	J= 9 I= 1		0.4116			0.0059
		2	0.6478			0.0113
		3	0.8926			0.0178
		4	1.1590			0.0258
		5	1.4574			0.0370
		6	1.7992			0.0517
		7	2.2163			0.0747
		8	2.7722			0.1197
		9	1.8779	8.1844	1.88E+1	0.2637
SUM OF MEAN:		7.7426				

TABLE III. Expected values of standard normal variates  
above the value 0.1 in groups of K=14

TRUNCATION VALUE = 0.1

		MEAN OF X	MEAN OF X**2	MEAN OF X**3	MEAN OF X**4	COVARIANCE
K=14	J= 1 I= 1	0.1763	0.0363	0.0089	2.58E-3	0.0053
	J= 2 I= 1		0.0499			0.0049
		0.2549	0.0753	0.0256	9.96E-3	0.0103
	J= 3 I= 1		0.0637			0.0047
			0.0951			0.0097
		0.3349	0.1271	0.0543	2.58E-2	0.0150
	J= 4 I= 1		0.0787			0.0045
			0.1166			0.0093
			0.1551			0.0142
	J= 5 I= 1	0.4209	0.1971	0.1019	5.77E-2	0.0200
			0.0937			0.0043
			0.1383			0.0090
			0.1834			0.0135
			0.2325			0.0190
	J= 6 I= 1	0.5074	0.2821	0.1706	1.11E-1	0.0247
			0.1099			0.0041
			0.1616			0.0086
			0.2139			0.0129
			0.2706			0.0180
			0.3278			0.0233
	J= 7 I= 1	0.6003	0.3901	0.2725	2.03E-1	0.0298
			0.1269			0.0039
			0.1860			0.0081
			0.2458			0.0121
			0.3107			0.0169
			0.3759			0.0218
			0.4469			0.0280
	J= 8 I= 1	0.6980	0.5223	0.4166	3.52E-1	0.0352
			0.1452			0.0036
			0.2124			0.0075
			0.2805			0.0114
			0.3542			0.0160
			0.4283			0.0205
			0.5088			0.0264
			0.5941			0.0332
	J= 9 I= 1	0.8037	0.6872	0.6221	5.93E-1	0.0414
			0.1651			0.0033
			0.2409			0.0070
			0.3180			0.0106
			0.4011			0.0149
			0.4846			0.0190
			0.5755			0.0247
			0.6716			0.0311
			0.7764			0.0388
	J=10 I= 1	0.9178	0.8907	0.9105	9.76E-1	0.0485
			0.1876			0.0031
			0.2733			0.0065
			0.3605			0.0100
			0.4544			0.0139
			0.5487			0.0176
			0.6514			0.0230
			0.7597			0.0291
			0.8777			0.0364
			1.0062			0.0456
	10	1.0468	1.1530	1.3322	1.61E+0	0.0572

TABLE III continued

J=11 I=	1	0.2140			0.0026
	2	0.3112			0.0056
	3	0.4103			0.0087
	4	0.5171			0.0124
	5	0.6243			0.0159
	6	0.7407			0.0210
	7	0.8636			0.0267
	8	0.9970			0.0333
	9	1.1422			0.0417
	10	1.3077			0.0525
	11	1.5067	1.9788	2.71E+0	0.0688
J=12 I=	1	0.2464			0.0024
	2	0.3579			0.0050
	3	0.4712			0.0076
	4	0.5938			0.0112
	5	0.7169			0.0145
	6	0.8502			0.0192
	7	0.9905			0.0244
	8	1.1429			0.0304
	9	1.3086			0.0381
	10	1.4970			0.0480
	11	1.7236			0.0637
	12	2.0053	3.0326	4.77E+0	0.0890
J=13 I=	1	0.2904			0.0024
	2	0.4212			0.0047
	3	0.5542			0.0071
	4	0.6980			0.0105
	5	0.8421			0.0134
	6	0.9984			0.0179
	7	1.1626			0.0225
	8	1.3412			0.0284
	9	1.5346			0.0355
	10	1.7542			0.0443
	11	2.0168			0.0580
	12	2.3427			0.0814
	13	2.7952	5.0017	9.34E+0	0.1269
J=14 I=	1	0.3659			0.0020
	2	0.5305			0.0042
	3	0.6979			0.0066
	4	0.8786			0.0098
	5	1.0598			0.0125
	6	1.2550			0.0160
	7	1.4608			0.0201
	8	1.6844			0.0254
	9	1.9260			0.0315
	10	2.1995			0.0387
	11	2.5256			0.0502
	12	2.9274			0.0698
	13	3.4803			0.1084
	14	4.4970	10.3215	2.49E+1	0.2358
SUM OF MEAN:	12.0424				

#### REFERENCES

Dixon, W.J. (1960), Simplified estimation from censored normal samples, Ann. Math. Stat. 31, 385-391.

Gupta, A.K. (1952), Estimation of the mean and standard deviation of a normal population from a censored sample, Biometrika 39, 260-273.

Sarhan, A.E. and Greenberg, B.G. (1956), Estimation of location and scale parameters by order statistics from singly and doubly censored samples. Part I. The normal distribution up to samples of size 10, Ann. Math. Stat. 27, 427-451.

Sarhan, A.E. and Greenberg, B.G. (1958), Estimation of location and scale parameters by order statistics from singly and doubly censored samples. Part II. Tables for the normal distribution for samples of size  $11 < N < 15$ , Ann. Math. Stat. 29, 79-105.

Saw, J.G. (1959), Estimation of the normal population parameters given a single censored sample. Biometrika 46, 150-159.



APPENDIX C

ESTIMATION OF THE PARAMETERS FOR SMALL DATA SETS OF  
LEFT CENSORED NORMAL AND LOG-NORMAL DATA

By:

Alan Gleit\*  
Versar Inc.  
P. O. Box 1549  
6850 Versar Center  
Springfield, Virginia 22151

\*This work was sponsored by the Air Force, Office of Scientific Research,  
under Contract F49620-82-C-0079.

# ABSTRACT

We study normal and lognormal data characterized by the dual problems of small sample size with several values reported as "smaller than the limit  $L$ ". In particular, we propose several estimators and report the results of a simulation.

Key words: Below detection limit; MLE; fill-in techniques; Type I censoring; environmental data analysis.

## INTRODUCTION

The reality of detection limits in the measurement of environmental phenomena is undeniable. Concentrations of pollutants are quite often too small to measure and are reported as "not detectable". For such measurements, we know only that the concentration lies below  $L$ , the detection limit.

Thus, the problem of address is one of censoring. In general, censoring means that observations at one or both extremes are not available. Our problem is equivalent to "left censoring"; life testing usually involves "right censoring", i.e., the largest values are not available. Two types of life censoring have received much attention. Type I occurs when the test is terminated at a specified time before all the items have failed; Type II occurs when the test is terminated at a particular failure. In Type I censoring the number of failures as well as the failure times are random variables. This, of course, makes Type I censoring far more complicated. Consequently, Type II methods have often been applied to Type I data with the hope that the bias is not appreciable. Our problem is analogous to Type I censoring since the number of measurements, say  $p$ , with concentrations below  $L$  is a random variable.

Environmental data is characterized not only by left censoring but also by small sample size. Required measurements for compliance purposes often are performed annually, quarterly, or, at most, monthly due to the expense or disruption caused by the testing. Studies of pilot plants or demonstration plants are often of such short duration that five to ten samples are all that are obtained. Thus, methods for estimating the parameters of environmental data using asymptotic or large-sample-size procedures are usually inapplicable.

In sum, environmental data usually has the following characteristics which make it difficult to analyze:

1. The data is left censored with a random number of data values.
2. The sample size is very small.

Further, environmental data quite often is well-modelled by the normal or the log-normal families. Consequently, the problem we address below is one of estimating the parameters of a normal or log-normal distribution when the data sets are characterized by (1) and (2) above.

The problem of estimating the parameters of left censored normal data has been extensively studied. Methods may be categorized as: (1) maximum likelihood estimators, (2) estimators based on linear combinations of order statistics, and (3) others. Maximum likelihood has been studied by, among others, Cohen (1950), Gupta (1952), and Harter and Moore (1966). Linear estimators have been studied by, among others, Gupta (1952), Sarhan and Greenberg (1956, 1958), Saw (1959), and Dixon (1960). Other methods include a method of moments suggested by Ipsen (1949) and the conservative estimator for the mean calculated by replacing all missing data by the truncation point (suggested by the U.S. Environmental Protection Agency).

All of the above techniques have drawbacks. The maximum likelihood procedures, though applicable to Type I data, are inefficient for small data sets and require numerical interpolation in extensive tables. The linear estimators are based on Type II censoring and so are biased for Type I data. Most estimation schema require extensive tables of coefficients. The moment estimator is extremely inefficient for small data sets while the conservative estimator is extremely biased.

Below we shall evaluate on simulated data various estimators, some new and some from the literature, to determine which (if any) are reasonable.

## SECTION 1. MAXIMUM LIKELIHOOD ESTIMATOR

In this section we recall the procedures from Gupta (1952). We assume  $0 < p < N$  values of our  $N$  samples lie below the (known) limit  $L$ . Thus we are given data

$$\{X_1, \dots, X_K, p \text{ values below } L\} \quad (1)$$

where we have taken

$$K = N - p. \quad (2)$$

We are asked to estimate the mean  $\mu$  and standard deviation  $\sigma$ . For our data (1) the likelihood function is given by:

$$\Phi((L-\mu)/\sigma) \prod_{i=1}^K \varphi((X_i - \mu)/\sigma) \quad (3)$$

where  $\Phi$ , respectively  $\varphi$ , is the cdf, respectively pdf, for the standard normal random variable. Taking logarithms yields

$$\log \text{likelihood} = p \log \Phi((L-\mu)/\sigma) - K \log \sigma - \frac{1}{2} \sum ((X-\mu)/\sigma)^2. \quad (4)$$

Maximizing the expression (4) yields the maximum likelihood

estimators (MLE). Setting the partial derivatives equal to zero yields

$$\mu = \frac{1}{K} \sum X - \frac{p\sigma}{K} \frac{\varphi((L-\mu)/\sigma)}{\Phi((L-\mu)/\sigma)} \quad (5)$$

$$\sigma^2 = \frac{1}{K} \sum (X-\mu)^2 - \frac{p\sigma}{K} (L-\mu) \frac{\varphi((L-\mu)/\sigma)}{\Phi((L-\mu)/\sigma)}. \quad (6)$$

Let

$$\bar{X} = \frac{1}{K} \sum X \quad (7)$$

$$S^2 = \frac{1}{K} \sum (X - \bar{X})^2. \quad (8)$$

The procedure to find  $\mu$  and  $\sigma$  is then as follows:

1. Calculate  $\bar{X}$ ,  $d$ ,  $S^2$ , and  $p/K$  from the data.
2. Calculate  $D$  using (19).
3. Find a value of  $a$  satisfying (13), (14), and (20). Find the corresponding value for  $z$ .
4. Then the estimates follow from (15) and (16):

$$\sigma^* = d/z \quad (21)$$

$$\mu^* = \bar{X} + (\sigma^{*2} - S^2) / d \quad (22)$$

In order to carry out this algorithm a table is needed giving the values of  $z$  for a given pair  $(D, p/K)$ . Tables can be found in Gupta (1952) and also below as Table 1. Note that when  $K=1$  we have  $S=d=0$  and so that above procedures do not produce useful results.

## SECTION 2. TRUNCATED MAXIMUM LIKELIHOOD ESTIMATOR

Our next technique is very easy to conceptualize: forget that data below  $L$  has been obtained and assume that the distribution of the remaining  $K$  data points are governed by the truncated normal distribution

$$g(X) = \varphi((X - \mu)/\sigma) / (1 - \Phi((L - \mu)/\sigma)). \quad (23)$$

The log likelihood for our  $K$  data points is

$$\log \text{likelihood} = -K \log(1 - \Phi(a)) - K \log \sigma - 1/2 \sum ((X - \mu)/\sigma)^2$$

where we again used (12):

$$a = (L - \mu)/\sigma.$$

Proceeding as in Section 1 we let

$$B(a) = \varphi(a)/(1 - \Phi(a)) \quad (24)$$

$$z_T(a) = -a + B(a). \quad (25)$$

Then

$$D = S^2 / (S^2 + d^2) \quad (26)$$

$$= (1 - az_T - z_T)^2 / (1 - az_T). \quad (27)$$

So we need to modify our previous algorithm in Step 3 to find  $z_T$  for a given value of  $D$ . Our Table 2 provides the necessary input.

### SECTION 3. CONDITIONAL FILL-IN TECHNIQUES

A general technique for dealing with missing values is to replace them with proxies. In this section we describe estimators using constants or using expected values of the censored values as proxies.

#### i. Fill-In with Constants

Various constants have been suggested as proxies for the data below L. The United States Environmental Protection Agency has a mandate to protect the human population from harmful pollutants. In doing this it usually errs on the side of conservatism. Thus, EPA often suggests that all censored values be replaced by the censoring value L to obtain the clearly most upward-biased, i.e., conservative, estimator for the mean pollutant levels. For pollutant concentrations the most liberal policy is the one that substitutes zero for the censored data: If I cannot measure it, it's not there. Those suggesting some sort of balance might use L/2. Let us suppose that we use the value C as a proxy. Then our "data" are

$$\{X_1, \dots, X_K, C, \dots, C\}.$$

Since we have all N values, we would use the usual estimators for the mean and variance:

$$\mu^* = \frac{1}{N} (\sum X + pC) \quad (28)$$

$$\sigma^{*2} = \frac{1}{N-1} (\sum X^2 + pC^2 - N\mu^{*2}) . \quad (29)$$

This procedure is very easy to use and is easily understood by the statistically non-sophisticated.



## 2. Fill-In with Random Order Statistics

As an alternative, we may elect to fill-in the censored data with seemingly more appropriate values: their expected values. To develop the formulas, we note the following.

### Proposition 1.

$$1. E(X | X \geq L) = \mu + \sigma \frac{\varphi((L-\mu)/\sigma)}{1 - \Phi((L-\mu)/\sigma)} \quad (30)$$

$$2. E(X | X \leq L) = \mu - \sigma \frac{\varphi((L-\mu)/\sigma)}{\Phi((L-\mu)/\sigma)} \quad (31)$$

$$3. E(X^2 | X \geq L) = \mu^2 + \sigma^2 + (L + \mu) \sigma \frac{\varphi((L-\mu)/\sigma)}{1 - \Phi((L-\mu)/\sigma)} \quad (32)$$

$$4. E(X^2 | X \leq L) = \mu^2 + \sigma^2 - (L + \mu) \sigma \frac{\varphi((L-\mu)/\sigma)}{\Phi((L-\mu)/\sigma)}. \quad (33)$$

Thus the expected values of the sum of the censored data and of the sums of squares are  $p$  times the right-hand-sides of (31) and (33), respectively. Hence,  $\mu^*$  and  $\sigma^*$  must satisfy

$$\mu^* = \frac{1}{N} [\sum X + p\mu^* - p\sigma^* \frac{\varphi((L - \mu^*)/\sigma^*)}{\Phi((L - \mu^*)/\sigma^*)}] \quad (34)$$

$$\sigma^{*2} = \frac{1}{N-1} [\sum X^2 + p\mu^{*2} + p\sigma^{*2} - (L + \mu^*) \sigma^* p \frac{\varphi((L - \mu^*)/\sigma^*)}{\Phi((L - \mu^*)/\sigma^*)} - N \mu^{*2}]. \quad (35)$$

Let us simplify these expressions. Recalling our previous definitions (7), (8), for  $\bar{X}$  and  $S^2$  and our previous notation (12), (13), for  $a$ ,  $A(a)$ , we find that (34) and (35) may be transformed to

$$\mu^* = \bar{X} - \frac{p\sigma^*}{K} A(a^*) \quad (36)$$

$$\sigma^{*2} = S^2 + \frac{p\sigma^*}{K-1} A(a^*) (\bar{X} - L) = S^2 + \frac{K}{K-1} (\bar{X} - \mu^*)(\bar{X} - L). \quad (37)$$

Except for the factor  $K/(K-1)$  in (37) these are identical to the MLE estimators (9) and (16)! Consequently, the MLE procedure is almost equivalent to filling-in the censored data with their conditional expectations. To numerically solve (36) and (37) set

$$\mu_0 = \sigma_0 = 1$$

and use the right-hand-sides of (36) and (37) to define  $\mu_{j+1}$ ,  $\sigma_{j+1}$  in terms of  $\mu_j$ ,  $\sigma_j$ .

#### SECTION 4. UNCONDITIONAL ESTIMATORS

The estimators developed in Sections 1, 2, and 3 are biased. The problem is that they estimate the parameters conditionally on the knowledge of  $p$ , the number of censored values. In this section we will readjust the estimators removing the bias due to conditioning.

##### 1. Fill-In with Constants

Recall that our  $N$  data values are

$$\{ X_1, \dots, X_K, C, \dots, C \}.$$

To compute the expected value of  $\mu^*$  and  $\sigma^*$  given by (28) and (29) we will use Proposition 1 and definition (24) for  $B$ .

##### Proposition 2.

$$1. \quad E(\Sigma X + pC|p) = K(\mu + \sigma B) + pC. \quad (38)$$

$$2. \quad E(\Sigma X^2 + pC^2|p) = K(\mu^2 + \sigma^2 + (L + \mu)\sigma B) + pC^2. \quad (39)$$

$$3. \quad E((\Sigma X + pC)^2|p) = K(K-1)(\mu + \sigma B)^2 + 2pKC(\mu + \sigma B) + p(p-1)C^2 \\ + K(\mu^2 + \sigma^2 + (L + \mu)\sigma B) + pC^2. \quad (40)$$

##### Corollary 3.

$$1. \quad E(\mu^*|p) = (K(\mu + \sigma B) + pC)/N \quad (41)$$

$$2. \quad E(\sigma^{*2}|p) = [K(\mu^2 + \sigma^2 + (L + \mu)\sigma B) + pC^2]/N - K(K-1)(\mu + \sigma B)^2/N(N-1) \\ - 2pKC(\mu + \sigma B)/N(N-1) - C^2p(p-1)/N(N-1). \quad (42)$$

So, given data, first compute  $\mu^*$  and  $\sigma^*$ . Using these values unbiased estimates  $\mu_0$ ,  $\sigma_0$  may be found by solving

$$\mu^* = (K(\mu_0 + \sigma_0 B_0) + pC)/N \quad (43)$$

$$\begin{aligned} \sigma^{*2} = & K[\mu_0^2 + \sigma_0^2 + (L + \mu_0)\sigma_0 B_0]/N - K(K-1)(\mu_0 + \sigma_0 B_0)^2/N(N-1) \\ & - 2pKC(\mu_0 + \sigma_0 B_0)/N(N-1) + KpC^2/N(N-1) \end{aligned} \quad (44)$$

where

$$B_0 = B((L - \mu_0)/\sigma_0). \quad (45)$$

Values for  $\mu_0$ ,  $\sigma_0$  satisfying these equations to any pre-set degree of accuracy may be easily obtained by use of a computer.

As an example, we initialize by

$$\mu_1 = \mu^*, \sigma_1 = \sigma^*$$

and then update by

$$B_j = B((L - \mu_j)/\sigma_j) \quad (46)$$

$$\mu_{j+1} = (N\mu^* - pC - \sigma_j K B_j)/K \quad (47)$$

$$\begin{aligned} \sigma_{j+1}^2 = & N\sigma^{*2}/K + (K-1)(\mu_{j+1} + \sigma_j B_j)^2/(N-1) + 2pC(\mu_{j+1} + \sigma_j B_j)/(N-1) \\ & - pC^2/(N-1) - \mu_{j+1}^2 - (L + \mu_{j+1})\sigma_j B_j. \end{aligned} \quad (48)$$

### 3. Fill-In with Random Order Statistics

We let  $Y_1 \leq Y_2 \leq \dots \leq Y_p \leq L$  be the (random) order statistics. We then use Proposition 1 and definitions (13) and (24) for A and B to obtain the following.

Proposition 4.

$$1. \quad E(\Sigma X + \Sigma Y|p) = N\mu + \sigma(KB - pA) \quad (49)$$

$$2. \quad E(\Sigma X^2 + \Sigma Y^2|p) = N(\mu^2 + \sigma^2) + \sigma(L + \mu)(KB - pA) \quad (50)$$

$$3. \quad E((\Sigma X + \Sigma Y)^2|p) = K(K-1)(\mu + \sigma B)^2 + 2pK(\mu + \sigma B)(\mu - \sigma A) + p(p-1)(\mu - \sigma A)^2 \\ + N(\mu^2 + \sigma^2) + \sigma(L + \mu)(KB - pA). \quad (51)$$

Using these results, we can compute the expected values of our estimators  $\mu^*$  and  $\sigma^*$  as given by (34) and (35).

Corollary 5.

$$1. \quad E(\mu^*|p) = \mu + \sigma(KB - pA)/N \quad (52)$$

$$2. \quad E(\sigma^{*2}|p) = [N(\mu^2 + \sigma^2) + \sigma(L + \mu)(KB - pA)]/N - K(K-1)(\mu + \sigma B)^2/N(N-1) \\ - 2pK(\mu + \sigma B)(\mu - \sigma A)/N(N-1) - p(p-1)(\mu - \sigma A)^2/N(N-1). \quad (53)$$

We may find unbiased estimators  $\mu_0$  and  $\sigma_0$  (to any degree of accuracy) for this case in a fashion similar to that of case 1 above.

One scheme puts

$$\mu_{j+1} = \mu^* - \sigma_j(KB_j - pA_j)/N \quad (54)$$

$$\sigma_{j+1}^2 = \sigma^{*2} + \frac{K(K-1)}{N(N-1)} (\mu_{j+1} + \sigma_j B_j)^2 + \frac{2pK}{N(N-1)} (\mu_{j+1} + \sigma_j B_j)(\mu_{j+1} - \sigma_j A_j) + \\ + \frac{p(p-1)(\mu_{j+1} - \sigma_j A_j)^2}{N(N-1)} - \frac{\sigma_j(L + \mu_{j+1})(KB_j - pA_j)}{N} - \mu_{j+1}^2. \quad (55)$$

### 3. Maximum Likelihood Estimators

We noted above that the maximum likelihood estimators (9) and (16) were virtually identical to the Fill-In with expected value estimators (36) and (37). Consequently, we can find the expected values for the MLEs.

Lemma 6.

$$E(S^2) = \sigma^2 + (L-\mu)\sigma B - \sigma^2 B^2. \quad (56)$$

Proposition 7.

$$1. \quad E(\mu^*|p) = \mu + \sigma(KB-pA)/N \quad (57)$$

$$\begin{aligned} 2. \quad E(\sigma^{*2}|p) &= E\left(\frac{K-1}{K}\sigma^{*2}(\text{E.V.}) + \frac{1}{K}S^2|p\right) \\ &= \frac{K-1}{K}\mu^2 + \sigma^2 + \frac{K-1}{KN}\sigma(L+\mu)(KB-pA) - \frac{(K-1)^2}{N(N-1)}(\mu+\sigma B)^2 - \\ &\quad - \frac{2p(K-1)}{N(N-1)}(\mu+\sigma B)(\mu-\sigma A) - \frac{p(p-1)(K-1)}{N(N-1)K}(\mu-\sigma A)^2 + \frac{(L-\mu)\sigma B}{K} - \frac{\sigma^2 B^2}{K}. \end{aligned} \quad (58)$$

We may find unbiased estimators  $\mu_0$  and  $\sigma_0$  (to any degree of accuracy) for this case in a fashion similar to case 1 above. One scheme puts

$$\mu_{j+1} = \mu^* - \sigma_j(KB_j - pA_j)/N \quad (59)$$

$$\begin{aligned} \sigma_{j+1}^2 &= \sigma^{*2} + \frac{(K-1)^2}{N(N-1)}(\mu_{j+1} + \sigma_j B_j)^2 + \frac{2p(K-1)}{N(N-1)}(\mu_{j+1} + \sigma_j B_j)(\mu_{j+1} - \sigma_j A_j) + \\ &\quad + \frac{p(p-1)(K-1)}{N(N-1)K}(\mu_{j+1} - \sigma_j A_j)^2 - \frac{K-1}{KN}\sigma_j(L+\mu_{j+1})(KB_j - pA_j) \\ &\quad + \frac{\sigma_j^2 B_j^2}{K} - \frac{K-1}{K}\mu_{j+1}^2. \end{aligned} \quad (60)$$

## SECTION 5. ORDER STATISTIC TECHNIQUES

Previous work on linear estimators are for Type II censoring, i.e., those with fixed sample sizes and not fixed censoring points. These have often been used in Type I situations with the hope that the resulting bias is small. We have investigated linear estimators for singly censored Type I samples elsewhere (Gleit 1983) and reported their very poor performance.

## SECTION 6. LOGNORMAL DATA

If our sample were from a lognormal distribution, then we could apply the techniques described above to the logarithms of the data to estimate  $\mu$  and  $\sigma$  for the resulting normal distribution. Then the parameters of the lognormal could be estimated by using the following facts:

$$\text{mean of lognormal} = \exp (\mu + \sigma^2/2)$$

$$\text{variance of lognormal} = \exp (2\mu + \sigma^2) (\exp (\sigma^2) - 1).$$



## SECTION 7. SIMULATED DATA

To evaluate the performance of our estimators we performed a simulation. By rescaling and changing origins, all the formulas depend on choosing two of the three parameters:  $\mu$ ,  $\sigma$ , and  $L$ . We normalized the simulated data to  $L=1$  and selected the following seven combinations for  $\mu$  and  $\sigma$ :

$$\mu = 0.67; \sigma = .2, .3$$

$$\mu = 1.00; \sigma = .1, .2, .3$$

$$\mu = 1.33; \sigma = .2, .3.$$

We selected  $N=5, 10$ , and  $15$  as representative small data set sizes.

Using a standard pseudo-random number generator and the Box-Muller transformation, we generated one million standard normal random variates. Using these variates, we generated 50,000 data sets for each of the twenty-one combinations of  $N$ ,  $\mu$ , and  $\sigma$ . These data sets were then artificially censored at the cutoff  $L=1$  and passed to the several estimators to "guess" values for  $\mu$  and  $\sigma$ . The data sets were then grouped by the value of  $p$ , the number of censored values,  $p=0, 1, \dots, N-1$ .

For each technique, each  $p$ , each  $N$ , and each  $\mu, \sigma$  combination we computed the mean and variance of the estimators for  $\mu$ , for  $\sigma$ , and for the mean and variance of the lognormal distribution whose logarithms follow the normal  $(\mu, \sigma)$  distribution.

Typical results are reported in Tables 3, 4, and 5 below. Table 3 reports the results for estimating the mean from data with  $N=5$ ,  $\mu=1.33$ , and  $\sigma=0.2$ . The sample sizes are large enough only for  $p=0, 1$ , and  $2$ . We see that the modified MLE and modified fill-in with constants routines failed to converge for  $p=2$  and did a fair job of converging for  $p=1$ . The truncation method had an unacceptably large variance. The MLE was very biased high; the modified version did not noticeably improve the

estimator. The expected value did a reasonable job while its modified form decreased the bias at the expense of added variance.

Table 4 also reports the results for estimating the mean but from data with  $N=10$ ,  $\mu=1.00$ , and  $\sigma=0.3$ . The sample sizes are large enough for all values of  $p$  from 1 to 9 (i.e., all values of interest except  $p=0$ ). Again the modified MLE and modified fill-in with constants routines did not converge very often. The truncation method again has large variance, the MLE is biased high, the expected value does very well, and the modified expected value does the best. Finally Table 5 reports the results for the mean for simulated lognormal data with mean 1.99,  $N=5$ , corresponding to  $\mu=0.67$  and  $\sigma=0.2$  for the underlying normal. The results are essentially the same as in Tables 3 and 4.

The results are very consistent throughout all the twenty-one cases for each of the four possible quantities estimated: normal mean and variance, lognormal mean and variance. The expected value estimator does a very good job while its modified form reduces the bias but increases the variance. These procedures converge just about all the time. The MLE is usually highly biased, has a large variance, and is not usable for the case  $K=1$  (i.e., only one data point). The modified MLE almost never converged; even when it did, it did a poor job. The truncation method always had an unacceptably large variance; it was also very biased for  $P/N$  large and not usable for  $K=1$ . Fill-in constants did not perform very well. For small  $p$ , fill-in with 0.5 did not do too badly; for large  $p$ , the estimator virtually agrees with the constant and so is of no value. The modified form almost never converges. Using the criteria of minimum square error, i.e.

$$\begin{aligned}\text{square error} &= E (\hat{\theta} - \theta)^2 \\ &= \text{Bias}^2 + \text{Variance},\end{aligned}$$

in general the modified expected value is best with fill-in by expected values coming in a close second.

### CONCLUSION

We have presented above several methods to estimate the mean and variance for a normal distribution based on censored from below data sets. Several are extremely simple, most require extensive computer calculations and some require extensive tables. Among these the modified fill-in by expected values (54) and (55) is our choice with fill-in by expected values (36) and (37) a close second. Though far more biased, this latter approach has lower variance.

# REFERENCES

1. Cohen, A.C. Jr., (1950), Estimating the mean and variance of normal populations from singly and doubly truncated samples, Ann. Math. Stat. 21, 557-569.
2. Dixon, W.J., (1960), Simplified estimation from censored normal samples, Ann. Math. Stat. 31, 385-391.
3. Gleit, A.S., (1983), Estimation of the normal population parameters by order statistics given a singly censored Type I sample.
4. Gupta, A.K., (1952), Estimation of the mean and standard deviation of a normal population from a censored sample, Biometrika 39, 260-273.
5. Harter, H.L. and Moore, A.H., (1966), Iterative maximum-likelihood estimation of the parameters of normal populations from singly and doubly censored samples, Biometrika 53, 205-213.
6. Ipsen, J., Jr., (1949), A practical method of estimating the mean and standard deviation of truncated normal distributions, Human Biology 21, 1-16.
7. Sarhan, A.E. and Greenberg, B.G., (1956), Estimation of location and scale parameters by order statistics from singly and doubly censored samples. Part I. The normal distribution up to samples of size 10. Ann. Math. Stat. 27, 427-451.
8. \_\_\_\_\_, (1958), Estimation of location and scale parameters by order statistics from singly and doubly censored samples. Part II. Tables for the normal distribution for samples of size  $11 < N < 15$ , Ann. Math. Stat. 29, 79-105.
9. Saw, J.G., (1959), Estimation of the normal population parameters given a single censored sample, Biometrika 46, 150-159.

Table 1. Values of  $z$  for given values of  $D$  and  $p/K$

		p/K					
		.0714	.1111	.1538	.2500	.3280	
D=	.0001	*	3.7478	3.0113	2.5570	2.0296	1.7010
	.0010	*	3.7229	2.9980	2.5555	2.0252	1.6983
	.0100	*	3.4965	2.8731	2.4776	1.9328	1.6721
	.0500	*	2.8093	2.4461	2.1627	1.8157	1.5640
	.0800	*	2.4784	2.2139	2.0097	1.7085	1.4921
	.1100	*	2.2294	2.0275	1.8640	1.6133	1.4251
	.1400	*	2.0321	1.8725	1.7400	1.5277	1.3631
	.1700	*	1.8699	1.7414	1.6311	1.4499	1.3051
	.2000	*	1.7329	1.6272	1.5345	1.3735	1.2505
	.2300	*	1.6140	1.5264	1.4478	1.3125	1.1991
	.2600	*	1.5107	1.4363	1.3690	1.2511	1.1504
	.2900	*	1.4182	1.3540	1.2969	1.1936	1.1036
	.3200	*	1.3340	1.2803	1.2301	1.1344	1.0593
	.3500	*	1.2584	1.2116	1.1680	1.0881	1.0165
	.3800	*	1.1864	1.1478	1.1096	1.0393	0.9753
	.4100	*	1.1232	1.0881	1.0540	0.9925	0.9353
	.4400	*	1.0625	1.0319	1.0027	0.9477	0.8965
	.4700	*	1.0052	0.9786	0.9530	0.9043	0.8587
	.5000	*	0.9511	0.9276	0.9054	0.8623	0.8216
	.5300	*	0.8994	0.8791	0.8594	0.8215	0.7853
	.5600	*	0.8499	0.8322	0.8150	0.7816	0.7494
	.5900	*	0.8020	0.7868	0.7717	0.7424	0.7140
	.6200	*	0.7557	0.7424	0.7294	0.7038	0.6780
	.6500	*	0.7105	0.6991	0.6876	0.6655	0.6435
	.6800	*	0.6662	0.6563	0.6466	0.6273	0.6081
	.7100	*	0.6223	0.6140	0.6050	0.5891	0.5726
	.7400	*	0.5787	0.5716	0.5640	0.5505	0.5364
	.7700	*	0.5348	0.5290	0.5232	0.5114	0.4994
	.8000	*	0.4905	0.4857	0.4809	0.4711	0.4612
	.8300	*	0.4450	0.4412	0.4373	0.4295	0.4214
	.8600	*	0.3976	0.3940	0.3916	0.3855	0.3793
	.8900	*	0.3473	0.3451	0.3429	0.3384	0.3335
	.9200	*	0.2920	0.2904	0.2889	0.2859	0.2826
	.9500	*	0.2277	0.2256	0.2260	0.2242	0.2222

Table 1 (cont.)

			p/K				
			.4265	.5000	.5600	.6750	1.000
D=	.0001	*	1.5769	1.4704	1.2952	1.1544	1.0930
	.0010	*	1.5747	1.4685	1.2940	1.1534	1.0922
	.0100	*	1.5535	1.4504	1.2814	1.1443	1.0844
	.0500	*	1.4646	1.3767	1.2277	1.1044	1.0495
	.0600	*	1.4037	1.3249	1.1843	1.0753	1.0243
	.1100	*	1.3467	1.2758	1.1523	1.0468	0.9993
	.1400	*	1.2930	1.2293	1.1167	1.0190	0.9746
	.1700	*	1.2424	1.1848	1.0820	0.9918	0.9503
	.2000	*	1.1945	1.1424	1.0483	0.9649	0.9262
	.2300	*	1.1480	1.1016	1.0155	0.9384	0.9022
	.2600	*	1.1049	1.0621	0.9836	0.9121	0.8787
	.2900	*	1.0629	1.0241	0.9522	0.8863	0.8550
	.3200	*	1.0224	0.9872	0.9216	0.8606	0.8317
	.3500	*	0.9833	0.9515	0.8913	0.8351	0.8083
	.3800	*	0.9455	0.9165	0.8616	0.8098	0.7850
	.4100	*	0.9082	0.8821	0.8322	0.7846	0.7610
	.4400	*	0.8722	0.8484	0.8030	0.7594	0.7381
	.4700	*	0.8368	0.8154	0.7741	0.7341	0.7147
	.5000	*	0.8020	0.7820	0.7452	0.7089	0.6908
	.5300	*	0.7677	0.7504	0.7165	0.6833	0.6670
	.5600	*	0.7337	0.7183	0.6876	0.6575	0.6423
	.5900	*	0.7000	0.6862	0.6587	0.6316	0.6180
	.6200	*	0.6665	0.6540	0.6295	0.6051	0.5929
	.6500	*	0.6327	0.6217	0.6000	0.5782	0.5673
	.6800	*	0.5987	0.5891	0.5699	0.5506	0.5408
	.7100	*	0.5642	0.5560	0.5393	0.5222	0.5136
	.7400	*	0.5295	0.5221	0.5077	0.4928	0.4855
	.7700	*	0.4934	0.4874	0.4749	0.4623	0.4557
	.8000	*	0.4563	0.4512	0.4409	0.4302	0.4246
	.8300	*	0.4174	0.4132	0.4048	0.3960	0.3913
	.8600	*	0.3760	0.3728	0.3661	0.3590	0.3552
	.8900	*	0.3312	0.3288	0.3237	0.3184	0.3154
	.9200	*	0.2809	0.2792	0.2756	0.2716	0.2700
	.9500	*	0.2213	0.2202	0.2181	0.2158	0.2147

Table 1 (cont.)

			p/K				
			1.1425	1.500	2.000	2.3333	2.750
D=	.0001	*	1.0365	0.9342	0.8431	0.8004	0.7585
	.0010	*	1.0357	0.9339	0.8428	0.8000	0.7586
	.0100	*	1.0290	0.9286	0.8385	0.7964	0.7557
	.0500	*	0.9985	0.9056	0.8213	0.7811	0.7423
	.1000	*	0.9765	0.8885	0.8079	0.7697	0.7313
	.1100	*	0.9545	0.8712	0.7943	0.7575	0.7215
	.1400	*	0.9325	0.8539	0.7810	0.7456	0.7110
	.1700	*	0.9106	0.8366	0.7671	0.7336	0.7002
	.2000	*	0.8892	0.8194	0.7532	0.7211	0.6874
	.2300	*	0.8676	0.8019	0.7393	0.7086	0.6782
	.2500	*	0.8463	0.7844	0.7251	0.6956	0.6670
	.2900	*	0.8245	0.7667	0.7106	0.6830	0.6555
	.3200	*	0.8034	0.7490	0.6959	0.6692	0.6435
	.3500	*	0.7819	0.7310	0.6811	0.6562	0.6312
	.3800	*	0.7606	0.7128	0.6658	0.6423	0.6184
	.4100	*	0.7396	0.6944	0.6504	0.6280	0.6057
	.4400	*	0.7171	0.6756	0.6345	0.6137	0.5922
	.4700	*	0.6953	0.6567	0.6183	0.5987	0.5787
	.5000	*	0.6731	0.6375	0.6014	0.5834	0.5644
	.5300	*	0.6506	0.6177	0.5843	0.5673	0.5498
	.5600	*	0.6276	0.5975	0.5667	0.5509	0.5346
	.5900	*	0.6045	0.5770	0.5487	0.5337	0.5190
	.6200	*	0.5806	0.5556	0.5295	0.5162	0.5024
	.6500	*	0.5563	0.5335	0.5100	0.4977	0.4850
	.6800	*	0.5309	0.5106	0.4894	0.4774	0.4665
	.7100	*	0.5049	0.4867	0.4679	0.4576	0.4473
	.7400	*	0.4775	0.4616	0.4446	0.4362	0.4265
	.7700	*	0.4492	0.4353	0.4207	0.4124	0.4046
	.8000	*	0.4196	0.4071	0.3947	0.3874	0.3806
	.8300	*	0.3867	0.3767	0.3664	0.3607	0.3547
	.8600	*	0.3517	0.3437	0.3351	0.3304	0.3256
	.8900	*	0.3127	0.3065	0.2997	0.2961	0.2923
	.9200	*	0.2680	0.2636	0.2586	0.2561	0.2533
	.9500	*	0.2134	0.2107	0.2076	0.2063	0.2046

Table 1 (cont.)

			p/K			
			4.000	6.500	9.000	14.00
D=	.0001	*	0.6779	0.5960	0.5520	0.5028
	.0010	*	0.6779	0.5956	0.5520	0.5028
	.0100	*	0.6755	0.5940	0.5507	0.5018
	.0500	*	0.6654	0.5872	0.5448	0.4976
	.0800	*	0.6582	0.5816	0.5403	0.4939
	.1100	*	0.6501	0.5760	0.5356	0.4902
	.1400	*	0.6424	0.5704	0.5306	0.4866
	.1700	*	0.6336	0.5644	0.5259	0.4824
	.2000	*	0.6257	0.5580	0.5210	0.4788
	.2300	*	0.6171	0.5516	0.5156	0.4741
	.2600	*	0.6081	0.5452	0.5102	0.4699
	.2900	*	0.5991	0.5380	0.5044	0.4652
	.3200	*	0.5896	0.5309	0.4981	0.4600
	.3500	*	0.5796	0.5233	0.4919	0.4553
	.3800	*	0.5697	0.5153	0.4852	0.4496
	.4100	*	0.5588	0.5070	0.4781	0.4439
	.4400	*	0.5480	0.4992	0.4710	0.4377
	.4700	*	0.5367	0.4901	0.4634	0.4315
	.5000	*	0.5250	0.4807	0.4550	0.4248
	.5300	*	0.5128	0.4708	0.4466	0.4176
	.5600	*	0.5002	0.4606	0.4373	0.4099
	.5900	*	0.4866	0.4497	0.4276	0.4017
	.6200	*	0.4722	0.4379	0.4175	0.3929
	.6500	*	0.4573	0.4255	0.4061	0.3832
	.6800	*	0.4416	0.4118	0.3943	0.3730
	.7100	*	0.4245	0.3975	0.3812	0.3613
	.7400	*	0.4061	0.3820	0.3672	0.3491
	.7700	*	0.3864	0.3647	0.3516	0.3354
	.8000	*	0.3649	0.3459	0.3343	0.3198
	.8300	*	0.3412	0.3248	0.3149	0.3021
	.8600	*	0.3144	0.3008	0.2921	0.2816
	.8900	*	0.2835	0.2728	0.2661	0.2572
	.9200	*	0.2470	0.2339	0.2339	0.2274
	.9500	*	0.2003	0.1955	0.1922	0.1881



Table 2. Values of  $z_T$  for given values of D

<u>D</u>		<u><math>z_T</math></u>
.0001	*	59.9950
.0010	*	31.6070
.0100	*	9.9449
.0500	*	4.3586
.0800	*	3.3536
.1100	*	2.8130
.1400	*	2.4074
.1700	*	2.0832
.2000	*	1.8216
.2300	*	1.5904
.2600	*	1.3846
.2900	*	1.1980
.3200	*	1.0263
.3500	*	0.8663
.3800	*	0.7150
.4100	*	0.5695
.4400	*	0.4255
.4700	*	0.2847
.5000	*	0.2388
.5300	*	0.2217
.5600	*	0.2108
.5900	*	0.2025
.6200	*	0.1955
.6500	*	0.1892
.6800	*	0.1833
.7100	*	0.1777
.7400	*	0.1721
.7700	*	0.1665
.8000	*	0.1606
.8300	*	0.1542
.8600	*	0.1471
.8900	*	0.1397
.9200	*	0.1283
.9500	*	0.1137

TABLE 3. ESTIMATES FOR MEAN. N=5, MEAN=1.33, STD. DEVIATION=0.20

	p=0			p=1			p=2		
	mean	variance	no convergence	mean	variance	no convergence	mean	variance	no convergence
Expected Value	1.35120	.00642		1.25593	.00477		1.11159	.00183	
Mod. expected value	1.31396	.01384	.04%	1.28700	.01050		1.12911	.00609	
MLE	1.35120	.00642		1.44727	.01252		1.58551	.02828	
Mod. MLE	1.33107	.00973		1.51544	.01573	23.2%			96.3%
Truncation	1.42278	.11079		1.43900	.17607		1.47777	.27182	
Fill with 0.0	1.35120	.00642	.04%	1.08215	.00522		0.81401	.00401	
Mod. fill with 0.0	1.31396	.01384		0.97043	.03953	28.5%			96.3%
Fill with 0.5	1.35120	.00642	.04%	1.18218	.00517		1.01402	.00401	
Mod. fill with 0.5	1.31396	.01384		1.15809	.02028	28.6%			96.3%
Fill with 1.0	1.35120	.00642	.04%	1.28218	.00516		1.21401	.00401	
Mod. fill with 1.0	1.31396	.01384		1.27902	.01730	28.7%			96.4%

Of the 50,000 samples, we had 38862 for p=0; 9960 for p=1; 1124 for p=2; 52 for p=3; and 2 for p=4.

TABLE 4. ESTIMATES FOR MEAN. N=10, MEAN=1.00, STD. DEVIATION=0.3

	p=1			p=2			p=3		
	mean	variance	no convergence	mean	variance	no convergence	mean	variance	no convergence
Expected Value	1.20218	.00289		1.16032	.00218		1.10810	.00157	
Mod. expected value	1.16217	.00583		1.14182	.00313		1.08973	.00269	
MLE	1.27125	.00493		1.31073	.00686		1.35184	.00953	
Mod. MLE	1.28125	.00650		1.38841	.01104	1.1%	1.45226	.01594	32.1%
Truncation	1.80790	1.22396		1.75373	1.08155		1.77130	1.16095	
Fill with 0.0	1.11413	.00315		0.99138	.00270		0.86593	.00221	
Mod. fill with 0.0	0.97105	.00731		0.81168	.00782	2.3%	0.65233	.00491	32.1%
Fill with 0.5	1.16414	.00315		1.09137	.00271		1.01593	.00220	
Mod. fill with 0.5	1.09815	.00590	0.2%	1.01804	.00607	2.3%	0.92327	.00406	32.1%
Fill with 1.0	1.21414	.00314		1.19137	.00271		1.16591	.00224	
Mod. fill with 1.0	1.13632	.01532	0.2%	1.14192	.01446	2.3%	1.11383	.01543	32.2%

Of the 50,000 samples, we had 42 for p=0; 485 for p=1; 2213 for p=2; 5878 for p=3; 10299 for p=4; 12487 for p=5; 10098 for p=6; 5891 for p=7; 2059 for p=8; 508 for p=9; and 40 for p=10

\*Note that the MLE techniques are only applicable for k≥2.

TABLE 4. ESTIMATES FOR MEAN. N=10, MEAN=1.00, STD. DEVIATION=0.3 (CONTINUED)

	p=4			p=5			p=6		
	mean	variance	no convergence	mean	variance	no convergence	mean	variance	no convergence
Expected Value	1.05208	.00122		0.99166	.00047		0.92088	.00135	
Mod. expected value	1.03701	.00263		0.98532	.00159		0.92508	.00370	
MLE	1.40746	.01479		1.46925	.02426		1.54523	.03954	
Mod. MLE			93.2%			100%			100%
Truncation	1.74879	1.16377		1.73972	1.19951		1.71624	1.25170	
Fill with 0.0	0.74375	.00195		0.61959	.00167		0.49539	.00127	
Mod. fill with 0.0			93.2%			100%			100%
Fill with 0.5	0.94374	.00197		0.86959	.00166		0.79538	.00130	
Mod. fill with 0.5			93.2%			100%			100%
Fill with 1.0	1.14377	.00190		1.11959	.00165		1.09538	.00129	
Mod. fill with 1.0			93.3%			100%			100%

TABLE 4. ESTIMATES FOR MEAN. N=10, MEAN=1.00, STD. DEVIATION=0.3 (CONTINUED)

	p=7			p=8			p=9		
	mean	variance	no convergence	mean	variance	no convergence	mean	variance	no convergence
Expected Value	0.81219	.00894		0.73750	.01956		0.77089	.00025	34.4%
Mod. expected value	0.81825	.01147		0.75141	.02306		0.79980	.00469	34.4%
MLE	1.64714	.07686		1.79783	.17551		*		
Mod. MLE			100%			100%	*		
Truncation	1.62858	1.09472		1.47612	.50257		*		
Fill with 0.0	0.37176	.00099		0.24833	.00066		0.12507	.00037	
Mod. fill with 0.0			100%			100%	0.32417	.08557	4.9%
Fill with 0.5	0.72176	.00098		0.64833	.00066		0.57507	.00037	
Mod. fill with 0.5			100%			100%	0.78603	.05250	12.6%
Fill with 1.0	1.07174	.00103		1.04833	.00065		1.02507	.00036	
Mod. fill with 1.0			100%			100%	1.17638	.02786	55.3%

TABLE 5. ESTIMATES FOR MEAN OF LOGNORMAL  
N=5, MEAN OF NORMAL=0.67, STD. DEVIATION OF NORMAL=0.2, MEAN OF LOGNORMAL=1.99

	p=3			p=4		
	mean	variance	no convergence	mean	variance	no convergence
Expected Value	2.47575	.00200		2.18701	.00104	3.4%
Mod. expected value	2.28093	.00854		2.12254	.00898	3.4%
MLE	3.35152	.19632		*		
Mod. MLE			99.6%	*		
Truncation	3.74070	37.30981		*		
Fill with 0.0	1.84645	.00531		1.39885	.00206	
Mod. fill with 0.0			99.8%	1.24028	.05850	15.3%
Fill with 0.5	2.19805	.00477		1.91930	.00230	
Mod. fill with 0.5			99.8%	1.87444	.02788	16.3%
Fill with 1.0	2.82056	.00433		2.76859	.00227	
Mod. fill with 1.0			99.8%	2.91499	.02276	18.7%

Of the 50,000 samples, we had 1 for p=1; 59 for p=2; 1001 for p=3; 10083 for p=4; and 38856 for p=5.

\*Note that the MLE techniques are only applicable for  $K \geq 2$ .

**END**

**FILMED**

**1-84**

**DTIC**